PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C07H 21/04

A1

(11) International Publication Number: WO 97/10365

(43) International Publication Date: 20 March 1997 (20.03.97)

(21) International Application Number: PCT/US96/14839

(22) International Filing Date: 13 September 1996 (13.09.96)

(30) Priority Data:
08/529,115 15 September 1995 (15.09.95) US

(71) Applicant *(for all designated States except US)*: AFFYMAX TECHNOLOGIES N.V. [NL/NL]; De Ruyderkade 62, Curaçao (AN).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: LOCKHART, David, J. [US/US]; 610 Mountain View Avenue, Mountain View, CA 94041 (US). BROWN, Eugene, L. [US/US]; 1388 Walnut Street, Newton Highlands, MA 02161 (US). WONG, Gordon [US/US]; 239 Clark Road, Brookline, MA 02146 (US). CHEE, Mark [AU/US]; 3199 Waverly Street, Palo Alto, CA 94306 (US). GINGERAS, Thomas, R. [US/US]; 528 Juniper Hill Drive, Encinitas, CA 92021 (US). MITTMANN, Michael, P. [US/US]; 2377 St. Francis Drive, Palo Alto, CA 94303 (US). LIPSHUTZ, Robert, J. [US/US]; 970 Palo Alto Avenue, Palo Alto, CA 94301 (US). FODOR, Stephen, P., A. [US/US]; 3863 Nathan Way, Palo Alto, CA 94303 (US). WANG, Chunwei

(74) Agents: HUNTER, Tom et al.; Townsend and Townsend and Crew L.L.P., 8th floor, Two Embarcadero Center, San Francisco, CA 94111-3834 (US).

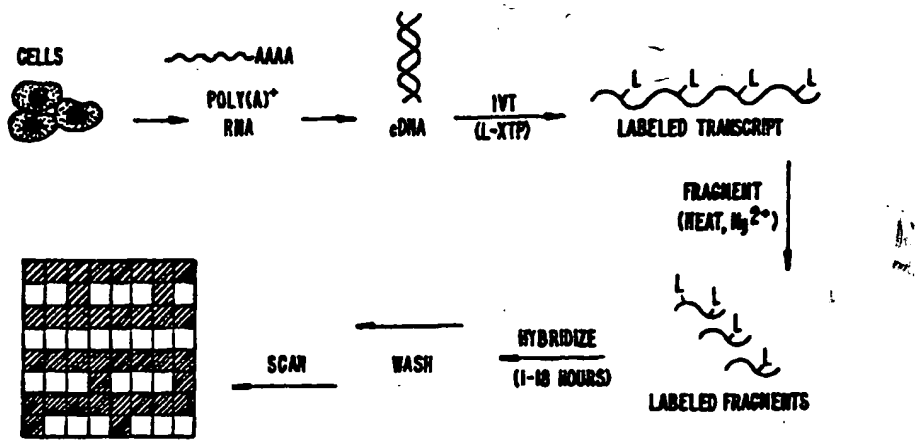
(81) Designated States: AU, CA, JP, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published

With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: EXPRESSION MONITORING BY HYBRIDIZATION TO HIGH DENSITY OLIGONUCLEOTIDE ARRAYS



(57) Abstract

This invention provides methods of monitoring the expression levels of a multiplicity of genes. The methods involve hybridizing a nucleic acid sample to a high density array of oligonucleotide probes where the high density array contains oligonucleotide probes complementary to subsequences of target nucleic acids in the nucleic acid sample. In one embodiment, the method involves providing a pool of target nucleic acids comprising RNA transcripts of one or more target genes, or nucleic acids derived from the RNA transcripts, hybridizing said pool of nucleic acids to an array of oligonucleotide probes immobilized on surface, where the array comprising more than 100 different oligonucleotides and each different oligonucleotide is localized in a predetermined region of the surface, the density of the different oligonucleotides is greater than about 60 different oligonucleotides per 1 cm^2, and the oligonucleotide probes are complementary to the RNA transcripts or nucleic acids derived from the RNA transcripts; and quantifying the hybridized nucleic acids in the array.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	**GB**	United Kingdom	**MW**	Malawi
AT	Austria	**GE**	Georgia	**MX**	Mexico
AU	Australia	**GN**	Guinea	**NE**	Niger
BB	Barbados	**GR**	Greece	**NL**	Netherlands
BE	Belgium	**HU**	Hungary	**NO**	Norway
BF	Burkina Faso	**IE**	Ireland	**NZ**	New Zealand
BG	Bulgaria	**IT**	Italy	**PL**	Poland
BJ	Benin	**JP**	Japan	**PT**	Portugal
BR	Brazil	**KE**	Kenya	**RO**	Romania
BY	Belarus	**KG**	Kyrgystan	**RU**	Russian Federation
CA	Canada	**KP**	Democratic People's Republic of Korea	**SD**	Sudan
CF	Central African Republic			**SE**	Sweden
CG	Congo	**KR**	Republic of Korea	**SG**	Singapore
CH	Switzerland	**KZ**	Kazakhstan	**SI**	Slovenia
CI	Côte d'Ivoire	**LI**	Liechtenstein	**SK**	Slovakia
CM	Cameroon	**LK**	Sri Lanka	**SN**	Senegal
CN	China	**LR**	Liberia	**SZ**	Swaziland
CS	Czechoslovakia	**LT**	Lithuania	**TD**	Chad
CZ	Czech Republic	**LU**	Luxembourg	**TG**	Togo
DE	Germany	**LV**	Latvia	**TJ**	Tajikistan
DK	Denmark	**MC**	Monaco	**TT**	Trinidad and Tobago
EE	Estonia	**MD**	Republic of Moldova	**UA**	Ukraine
ES	Spain	**MG**	Madagascar	**UG**	Uganda
FI	Finland	**ML**	Mali	**US**	United States of America
FR	France	**MN**	Mongolia	**UZ**	Uzbekistan
GA	Gabon	**MR**	Mauritania	**VN**	Viet Nam

EXPRESSION MONITORING BY HYBRIDIZATION TO HIGH DENSITY OLIGONUCLEOTIDE ARRAYS

CROSS REFERENCE TO RELATED APPLICATIONS

This is a continuation-in-part of U.S.S.N. 08/529,115 filed on September 15, 1995 which is herein incorporated by reference for all purposes.

BACKGROUND OF THE INVENTION

Many disease states are characterized by differences in the expression levels of various genes either through changes in the copy number of the genetic DNA or through changes in levels of transcription (*e.g.* through control of initiation, provision of RNA precursors, RNA processing, *etc.*) of particular genes. For example, losses and gains of genetic material play an important role in malignant transformation and progression. These gains and losses are thought to be "driven" by at least two kinds of genes. Oncogenes are positive regulators of tumorgenesis, while tumor suppressor genes are negative regulators of tumorgenesis (Marshall, *Cell*, 64: 313-326 (1991); Weinberg, Science, 254: 1138-1146 (1991)). Therefore, one mechanism of activating unregulated growth is to increase the number of genes coding for oncogene proteins or to increase the level of expression of these oncogenes (*e.g.* in response to cellular or environmental changes), and another is to lose genetic material or to decrease the level of expression of genes that code for tumor suppressors. This model is supported by the losses and gains of genetic material associated with glioma progression (Mikkelson *et al.* J. Cellular Biochm. 46: 3-8 (1991)). Thus, changes in the expression (transcription) levels of

Similarly, as hybridization efficiency and cross-reactivity varies with the particular subsequence (region) of a gene being probed it is difficult to obtain an accurate and reliable measure of gene expression with one, or even a few, probes to the target gene.

The development of VLSIPS™ technology provided methods for synthesizing arrays of many different oligonucleotide probes that occupy a very small surface area. *See* U.S. Patent No. 5,143,854 and PCT patent publication No. WO 90/15070. U.S. Patent application Serial No. 082,937, filed June 25, 1993, describes methods for making arrays of oligonucleotide probes that can be used to provide the complete sequence of a target nucleic acid and to detect the presence of a nucleic acid containing a specific nucleotide sequence.

Prior to the present invention, however, it was unknown that high density oligonucleotide arrays could be used to reliably monitor message levels of a multiplicity of preselected genes in the presence of a large abundance of other (non-target) nucleic acids (*e.g.*, in a cDNA library, DNA reverse transcribed from an mRNA, mRNA used directly or amplified, or polymerized from a DNA template). In addition, the prior art provided no rapid and effective method for identifying a set of oligonucleotide probes that maximize specific hybridization efficacy while minimizing cross-reactivity nor of using hybridization patterns (in particular hybridization patterns of a multiplicity of oligonucleotide probes in which multiple oligonucleotide probes are directed to each target nucleic acid) for quantification of target nucleic acid concentrations.

Summary of the Invention

The present invention is premised, in part, on the discovery that microfabricated arrays of large numbers of different oligonucleotide probes (DNA chips) may effectively be used to not only detect the presence or absence of target nucleic acid sequences, but to quantify the relative abundance of the target sequences in a complex nucleic acid pool. In addition, it was also a surprising discovery that relatively short oligonucleotide probes (*e.g.*, 20 mer) are sufficiently specific to allow quantitation of gene expression in complex mixtures of nucleic acids particularly when provided as in high density oligonucleotide probe arrays.

In a preferred embodiment, the array of oligonucleotide probes is a high density array comprising greater than about 100, preferably greater than about 1,000 more preferably greater than about 16,000 and most preferably greater than about 65,000 or 250,000 or even 1,000,000 different oligonucleotide probes. Such high density arrays comprise a probe density of generally greater than about 60, more generally greater than about 100, most generally greater than about 600, often greater than about 1000, more often greater than about 5,000, most often greater than about 10,000, preferably greater than about 40,000 more preferably greater than about 100,000, and most preferably greater than about 400,000 different oligonucleotide probes per cm^2 (where different oligonucleotides refers to oligonucleotides having different sequences). The oligonucleotide probes range from about 5 to about 50 nucleotides, preferably from about 5 to about 45 nucleotides, still more preferably from about 10 to about 40 nucleotides and most preferably from about 15 to about 40 nucleotides in length. Particularly preferred arrays contain probes ranging from about 20 to about 25 oligonucleotides in length. The array may comprise more than 10, preferably more than 50, more preferably more than 100, and most preferably more than 1000 oligonucleotide probes specific for each target gene. In a preferred embodiment, the array comprises at least 10 different oligonucleotide probes for each gene. In another preferred embodiment, the array 20 or fewer oligonucleotides complementary each gene. Although a planar array surface is preferred, the array may be fabricated on a surface of virtually any shape or even a multiplicity of surfaces.

The array may further comprise mismatch control probes. Where such mismatch controls are present, the quantifying step may comprise calculating the difference in hybridization signal intensity between each of the oligonucleotide probes and its corresponding mismatch control probe. The quantifying may further comprise calculating the average difference in hybridization signal intensity between each of the oligonucleotide probes and its corresponding mismatch control probe for each gene.

The probes present in the high density array can be oligonucleotide probes selected according to selection and optimization methods described below. Alternatively, non-optimal probes may be included in the array, but the probes used for

Hybridization may include subsequent washes at progressively increasing stringency until a desired level of hybridization specificity is reached.

Quantification of the hybridization signal can be by any means known to one of skill in the art. However, in a particularly preferred embodiment, quantification is achieved by use of a confocal fluorescence microscope. Data is preferably evaluated by calculating the difference in hybridization signal intensity between each oligonucleotide probe and its corresponding mismatch control probe. It is particularly preferred that this difference be calculated and evaluated for each gene. Particularly preferred analytical methods are provided herein.

The pool of target nucleic acids can be the total polyA$^+$ mRNA isolated from a biological sample, or cDNA made by reverse transcription of the RNA or second strand cDNA or RNA transcribed from the double stranded cDNA intermediate. Alternatively, the pool of target nucleic acids can be treated to reduce the complexity of the sample and thereby reduce the background signal obtained in hybridization. In one approach, a pool of mRNAs, derived from a biological sample, is hybridized with a pool of oligonucleotides comprising the oligonucleotide probes present in the high density array. The pool of hybridized nucleic acids is then treated with RNase A which digests the single stranded regions. The remaining double stranded hybridization complexes are then denatured and the oligonucleotide probes are removed, leaving a pool of mRNAs enhanced for those mRNAs complementary to the oligonucleotide probes in the high density array.

In another approach to background reduction, a pool of mRNAs derived from a biological sample is hybridized with paired target specific oligonucleotides where the paired target specific oligonucleotides are complementary to regions flanking subsequences of the mRNAs complementary to the oligonucleotide probes in the high density array. The pool of hybridized nucleic acids is treated with RNase H which digests the hybridized (double stranded) nucleic acid sequences. The remaining single stranded nucleic acid sequences which have a length about equivalent to the region flanked by the paired target specific oligonucleotides are then isolated (*e.g.* by electrophoresis) and used as the pool of nucleic acids for monitoring gene expression.

drugs to identify those that have an effect on particular gene(s). This involves providing a pool of target nucleic acids from one or more cells contacted with the drug or drugs and hybridizing that pool to any of the high density oligonucleotide arrays described herein. The expression levels of the genes targeted by the probes in the array are determined and compared to expression levels of genes from "control" cells not exposed to the drug or drugs. The genes that are overexpressed or underexpressed in response to the drug or drugs are identified or conversely the drug or drugs that alter expression of one or more genes are identified.

In still yet another embodiment, this invention provide for a composition comprising any of the high density oligonucleotide arrays disclosed herein where the oligonucleotide probes are specifically hybridized to one or more fluorescently labeled nucleic acids (which are the transcription products of genes or derived from those transcription products) thereby forming a fluorescent array in which the fluorescence of the array is indicative of the transcription levels of the multiplicity of genes. One of skill will appreciate that such a hybridized array may be used as a reference, control, or standard (*e.g.*, provided in a kit) or may itself be a diagnostic array indicating the expression levels of a multiplicity of genes in a sample.

This invention also provides kits for simultaneously monitoring expression levels of a multiplicity of genes. The kits include an array of immobilized oligonucleotide probes complementary to subsequences of the multiplicity of target genes, as described herein. The kit may also include instructions describing the use of the array for detection and/or quantification of expression levels of the multiplicity of genes. The kit may additionally include one or more of the following: buffers, hybridization mix, wash and read solutions, labels, labeling reagents (enzymes *etc.*), "control" nucleic acids, software for probe selection, array reading or data analysis and any of the other materials or reagents described herein for the practice of the claimed methods.

In another embodiment, this invention provides for a method of selecting a set of oligonucleotide probes, that specifically bind to a target nucleic acid (*e.g.*, a gene or genes whose expression is to be monitored or nucleic acids derived from the gene or its transcribed mRNA). The method involves providing a high density array of

hybridization intensities of the perfect match and mismatch probes of each pair are calculated.

Additionally, the invention provides a computer-implemented method for monitoring expression of genes comprising the steps of: receiving input of a nucleic acid sequence constituting a gene; generating a set of probes that are perfectly complementary to the gene; and identifying a subset of probes, including less than all of the probes in the set, for monitoring the expression of the gene. Each probe of the set may be analyzed by criteria that specify characteristics indicative of low hybridization or high cross hybridization. The criteria may include if occurrences of a specific nucleotide in a probe crosses a threshold value, if the number of a specific nucleotide that repeats sequentially in a probe crosses a threshold value, if the length of a palindrome in a probe crosses a threshold value, and the like.

Definitions.

The phrase "massively parallel screening" refers to the simultaneous screening of at least about 100, preferably about 1000, more preferably about 10,000 and most preferably about 1,000,000 different nucleic acid hybridizations.

The terms "nucleic acid" or "nucleic acid molecule" refer to a deoxyribonucleotide or ribonucleotide polymer in either single-or double-stranded form, and unless otherwise limited, would encompass known analogs of natural nucleotides that can function in a similar manner as naturally occurring nucleotides.

An oligonucleotide is a single-stranded nucleic acid ranging in length from 2 to about 500 bases.

As used herein a "probe" is defined as an oligonucleotide capable of binding to a target nucleic acid of complementary sequence through one or more types of chemical bonds, usually through complementary base pairing, usually through hydrogen bond formation. As used herein, an oligonucleotide probe may include natural (*i.e.* A, G, C, or T) or modified bases (7-deazaguanosine, inosine, *etc.*). In addition, the bases in oligonucleotide probe may be joined by a linkage other than a phosphodiester bond, so long as it does not interfere with hybridization. Thus, oligonucleotide probes may be

complementary to the target sequence hybridize to the target sequence at equilibrium. (As the target sequences are generally present in excess, at Tm, 50% of the probes are occupied at equilibrium). Typically, stringent conditions will be those in which the salt concentration is at least about 0.01 to 1.0 M Na ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30°C for short probes (*e.g.*, 10 to 50 nucleotides). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide.

The term "perfect match probe" refers to a probe that has a sequence that is perfectly complementary to a particular target sequence. The test probe is typically perfectly complementary to a portion (subsequence) of the target sequence. The perfect match (PM) probe can be a "test probe", a "normalization control" probe, an expression level control probe and the like. A perfect match control or perfect match probe is, however, distinguished from a "mismatch control" or "mismatch probe."

The term "mismatch control" or "mismatch probe" refer to probes whose sequence is deliberately selected not to be perfectly complementary to a particular target sequence. For each mismatch (MM) control in a high-density array there typically exists a corresponding perfect match (PM) probe that is perfectly complementary to the same particular target sequence. The mismatch may comprise one or more bases. While the mismatch(s) may be locates anywhere in the mismatch probe, terminal mismatches are less desirable as a terminal mismatch is less likely to prevent hybridization of the target sequence. In a particularly preferred embodiment, the mismatch is located at or near the center of the probe such that the mismatch is most likely to destabilize the duplex with the target sequence under the test hybridization conditions.

The terms "background" or "background signal intensity" refer to hybridization signals resulting from non-specific binding, or other interactions, between the labeled target nucleic acids and components of the oligonucleotide array (*e.g.*, the oligonucleotide probes, control probes, the array substrate, *etc.*). Background signals may also be produced by intrinsic fluorescence of the array components themselves. A single background signal can be calculated for the entire array, or a different background signal may be calculated for each target nucleic acid. In a preferred embodiment, background is calculated as the average hybridization signal intensity for the lowest 5%

Methods of alignment of sequences for comparison are well known in the art. Optimal alignment of sequences for comparison may be conducted by the local homology algorithm of Smith and Waterman, *Adv. Appl. Math.* 2: 482 (1981), by the homology alignment algorithm of Needleman and Wunsch *J. Mol. Biol.* 48: 443 (1970), by the search for similarity method of Pearson and Lipman, *Proc. Natl. Acad. Sci. USA* 85: 2444 (1988), by computerized implementations of these algorithms (including, but not limited to CLUSTAL in the PC/Gene program by Intelligenetics, Moutain View, California, GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group (GCG), 575 Science Dr., Madison, Wisconsin, USA), or by inspection. In particular, methods for aligning sequences using the CLUSTAL program are well described by Higgins and Sharp in *Gene*, 73: 237-244 (1988) and in *CABIOS* 5: 151-153 (1989)).

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 shows a schematic of expression monitoring using oligonucleotide arrays. Extracted poly $(A)^+$ RNA is converted to cDNA, which is then transcribed in the presence of labeled ribonucleotide triphosphates. L is either biotin or a dye such as fluorescein. RNA is fragmented with heat in the presence of magnesium ions. Hybridizations are carried out in a flow cell that contains the two-dimensional DNA probe arrays. Following a brief washing step to remove unhybridized RNA, the arrays are scanned using a scanning confocal microscope. Alternatives in which cellular mRNA is directly labeled without a cDNA intermediate are described in the Examples. Image analysis software converts the scanned array images into text files in which the observed intensities at specific physical locations are associated with particular probe sequences.

Fig. 2A shows a fluorescent image of a high density array containing over 16,000 different oligonucleotide probes. The image was obtained following hybridization (15 hours at 40°C) of biotin-labeled randomly fragmented sense RNA transcribed from the murine B cell (Tl0) cDNA library, and spiked at the level of 1:3,000 (50 pM equivalent to about 100 copies per cell) with 13 specific RNA targets. The brightness at any location is indicative of the amount of labeled RNA hybridized to the particular oligonucleotide probe. Fig. 2B shows a small portion of the array (the boxed region of Fig. 2A) containing probes

IL-4, IL-6, and IL-12p40 were not detected above the level of approximately 1:200,000 in these experiments. The error bars reflect the estimated uncertainty (25 percent) in the level for a given RNA relative to the level for the same RNA at a different time point. The relative uncertainty estimate was based on the results of repeated spiking experiments, and on repeated measurements of IL-10, β-actin and GAPDH RNAs in preparations from both T10 and 2D6 cells (unstimulated). The uncertainty in the absolute frequencies includes message-to-message differences in the hybridization efficiency as well as differences in the mRNA isolation, cDNA synthesis, and RNA synthesis and labeling steps. The uncertainty in the absolute frequencies is estimated to be a factor of three.

Fig. 5 shows a fluorescence image of an array containing over 63,000 different oligonucleotide probes for 118 genes. The image was obtained following overnight hybridization of a labeled murine B cell RNA sample. Each square synthesis region is 50 x 50 μm and contains 107 to 108 copies of a specific oligonucleotide. The array was scanned at a resolution of 7.5 μm in approximately 15 minutes. The bright rows indicate RNAs present at high levels. Lower level RNAs were unambiguously detected based on quantitative evaluation of the hybridization patterns. A total of 21 murine RNAs were detected at levels ranging from approximately 1:300,000 to 1:100. The cross in the center, the checkerboard in the corners, and the MUR-1 region at the top contain probes complementary to a labeled control oligonucleotide that was added to all samples.

Fig. 6 shows an example of a computer system used to execute the software of an embodiment of the present invention.

Fig. 7 shows a system block diagram of a typical computer system used to execute the software of an embodiment of the present invention.

Fig. 8 shows the high level flow of a process of monitoring the expression of a gene by comparing hybridization intensities of pairs of perfect match and mismatch probes.

Fig. 9 shows the flow of a process of determining if a gene is expressed utilizing a decision matrix.

Figs. 10A and 10B show the flow of a process of determining the expression of a gene by comparing baseline scan data and experimental scan data.

gene expression levels. Of course arrays containing longer oligonucleotides, as described herein, are also suitable.

A) Advantages of Oligonucleotide Arrays

In one preferred embodiment, the high density arrays used in the methods of this invention comprise chemically synthesized oligonucleotides. The use of chemically synthesized oligonucleotide arrays, as opposed to, for example, blotted arrays of genomic clones, restriction fragments, oligonucleotides, and the like, offers numerous advantages. These advantages generally fall into four categories:

1) Efficiency of production;
2) Reduced intra- and inter-array variability;
3) Increased information content; and
4) Higher signal to noise ratio (improved sensitivity).

1) Efficiency of production.

In a preferred embodiment, the arrays are synthesized using methods of spatially addressed parallel synthesis (*see, e.g.*, Section V, below). The oligonucleotides are synthesized chemically in a highly parallel fashion covalently attached to the array surface. This allows extremely efficient array production. For example, arrays containing tens (or even hundreds) of thousands of specifically selected 20 mer oligonucleotides are synthesized in fewer than 80 synthesis cycles. The arrays are designed and synthesized based on sequence information alone. Thus, unlike blotting methods, the array preparation requires no handling of biological materials. There is no need for cloning steps, nucleic acid amplifications, cataloging of clones or amplification products, and the like. The preferred chemical synthesis of expression monitoring arrays in this invention is thus more efficient blotting methods and permits the production of highly reproducible high-density arrays with relatively little labor and expense.

2) Reduced intra- and inter-array variability.

The use of chemically synthesized high-density oligonucleotide arrays in the methods of this invention improves intra- and inter-array variability. The

advantages not found with other methods. For example, the use of large numbers of different probes that specifically bind to the transcription product of a particular target gene provides a high degree of redundancy and internal control that permits optimization of probe sets for effective detection of particular target genes and minimizes the possibility of errors due to cross-reactivity with other nucleic acid species.

Apparently suitable probes often prove ineffective for expression monitoring by hybridization. For example, certain subsequences of a particular target gene may be found in other regions of the genome and probes directed to these subsequences will cross-hybridize with the other regions and not provide a signal that is a meaningful measure of the expression level of the target gene. Even probes that show little cross reactivity may be unsuitable because they generally show poor hybridization due to the formation of structures that prevent effective hybridization. Finally, in sets with large numbers of probes, it is difficult to identify hybridization conditions that are optimal for all the probes in a set. Because of the high degree of redundancy provided by the large number of probes for each target gene, it is possible to eliminate those probes that function poorly under a given set of hybridization conditions and still retain enough probes to a particular target gene to provide an extremely sensitive and reliable measure of the expression level (transcription level) of that gene.

In addition, the use of large numbers of different probes to each target gene makes it possible to monitor expression of families of closely-related nucleic acids. The probes may be selected to hybridize both with subsequences that are conserved across the family and with subsequences that differ in the different nucleic acids in the family. Thus, hybridization with such arrays permits simultaneous monitoring of the various members of a gene family even where the various genes are approximately the same size and have high levels of homology. Such measurements are difficult or impossible with traditional hybridization methods.

Because the high density arrays contain such a large number of probes it is possible to provide numerous controls including, for example, controls for variations or mutations in a particular gene, controls for overall hybridization conditions, controls for sample preparation conditions, controls for metabolic activity of the cell from which

to produce a given signal thereby dramatically improving the signal to noise ratio. Consequently the methods of this invention permit detection of only a few copies of a nucleic acid in extremely complex nucleic acid mixtures.

B) Preferred High Density Arrays

Preferred high density arrays of this invention comprise greater than about 100, preferably greater than about 1000, more preferably greater than about 16,000 and most preferably greater than about 65,000 or 250,000 or even greater than about 1,000,000 different oligonucleotide probes. The oligonucleotide probes range from about 5 to about 50 or about 5 to about 45 nucleotides, more preferably from about 10 to about 40 nucleotides and most preferably from about 15 to about 40 nucleotides in length. In particular preferred embodiments, the oligonucleotide probes are 20 or 25 nucleotides in length. It was a discovery of this invention that relatively short oligonucleotide probes sufficient to specifically hybridize to and distinguish target sequences. Thus in one preferred embodiment, the oligonucleotide probes are less than 50 nucleotides in length, generally less than 46 nucleotides, more generally less than 41 nucleotides, most generally less than 36 nucleotides, preferably less than 31 nucleotides, more preferably less than 26 nucleotides,and most preferably less than 21 nucleotides in length. The probes can also be less than 16 nucleotides or less than even 11 nucleotides in length.

The location and sequence of each different oligonucleotide probe sequence in the array is known. Moreover, the large number of different probes occupies a relatively small area providing a high density array having a probe density of generally greater than about 60, more generally greater than about 100, most generally greater than about 600, often greater than about 1000, more often greater than about 5,000, most often greater than about 10,000, preferably greater than about 40,000 more preferably greater than about 100,000, and most preferably greater than about 400,000 different oligonucleotide probes per cm^2. The small surface area of the array (often less than about 10 cm^2, preferably less than about 5 cm^2 more preferably less than about 2 cm^2, and most preferably less than about 1.6 cm^2) permits extremely uniform

In another preferred embodiment, the methods of this invention are used to monitor expression of various genes in response to defined stimuli, such as a drug. The methods are particularly advantageous because they permit simultaneous monitoring of the expression of thousands of genes. This is especially useful in drug research if the end point description is a complex one, not simply asking if one particular gene is overexpressed or underexpressed. Thus, where a disease state or the mode of action of a drug is not well characterized, the methods of this invention allow rapid determination of the particularly relevant genes.

As indicated above, the materials and methods of this invention are typically used to monitor the expression of a multiplicity of different genes simultaneously. Thus, in one embodiment, the invention provide for simultaneous monitoring of at least about 10, preferably at least about 100, more preferably at least about 1000, still more preferably at least about 10,000, and most preferably at least about 100,000 different genes.

The expression monitoring methods of this invention can also be used for gene discovery. Many genes that have been discovered to date have been classified into families based on commonality of the sequences. Because of the extremely large number of probes it is possible to place in the high density array, it is possible to include oligonucleotide probes representing known or parts of known members from every gene class. In utilizing such a "chip" (high density array) genes that are already known would give a positive signal at loci containing both variable and common regions. For unknown genes, only the common regions of the gene family would give a positive signal. The result would indicate the possibility of a newly discovered gene.

The expression monitoring methods of this invention also allow the development of "dynamic" gene databases. The Human Genome Project and commercial sequencing projects have generated large static databases which list thousands of sequences without regard to function or genetic interaction. Expression analysis using the methods of this invention produces "dynamic" databases that define a gene's function and its interactions with other genes. Without the ability to monitor the expression of large numbers of genes simultaneously ,however, the work of creating such a database is enormous. The tedious nature of using DNA sequence analysis for

transcript or a subsequence thereof has ultimately served as a template. Thus, a cDNA reverse transcribed from an mRNA, an RNA transcribed from that cDNA, a DNA amplified from the cDNA, an RNA transcribed from the amplified DNA, *etc.*, are all derived from the mRNA transcript and detection of such derived products is indicative of the presence and/or abundance of the original transcript in a sample. Thus, suitable samples include, but are not limited to, mRNA transcripts of the gene or genes, cDNA reverse transcribed from the mRNA, cRNA transcribed from the cDNA, DNA amplified from the genes, RNA transcribed from amplified DNA, and the like.

In a particularly preferred embodiment, where it is desired to quantify the transcription level (and thereby expression) of a one or more genes in a sample, the nucleic acid sample is one in which the concentration of the mRNA transcript(s) of the gene or genes, or the concentration of the nucleic acids derived from the mRNA transcript(s), is proportional to the transcription level (and therefore expression level) of that gene. Similarly, it is preferred that the hybridization signal intensity be proportional to the amount of hybridized nucleic acid. While it is preferred that the proportionality be relatively strict (*e.g.*, a doubling in transcription rate results in a doubling in mRNA transcript in the sample nucleic acid pool and a doubling in hybridization signal), one of skill will appreciate that the proportionality can be more relaxed and even non-linear. Thus, for example, an assay where a 5 fold difference in concentration of the target mRNA results in a 3 to 6 fold difference in hybridization intensity is sufficient for most purposes. Where more precise quantification is required appropriate controls can be run to correct for variations introduced in sample preparation and hybridization as described herein. In addition, serial dilutions of "standard" target mRNAs can be used to prepare calibration curves according to methods well known to those of skill in the art. Of course, where simple detection of the presence or absence of a transcript is desired, no elaborate control or calibration is required.

In the simplest embodiment, such a nucleic acid sample is the total mRNA isolated from a biological sample. The term "biological sample", as used herein, refers to a sample obtained from an organism or from components (*e.g.*, cells) of an organism. The sample may be of any biological tissue or fluid. Frequently the sample will be a "clinical sample" which is a sample derived from a patient. Such samples include, but

then include probes specific to the internal standard for quantification of the amplified nucleic acid.

One preferred internal standard is a synthetic AW106 cRNA. The AW106 cRNA is combined with RNA isolated from the sample according to standard techniques known to those of skill in the art. The RNA is then reverse transcribed using a reverse transcriptase to provide copy DNA. The cDNA sequences are then amplified (*e.g.*, by PCR) using labeled primers. The amplification products are separated, typically by electrophoresis, and the amount of radioactivity (proportional to the amount of amplified product) is determined. The amount of mRNA in the sample is then calculated by comparison with the signal produced by the known AW106 RNA standard. Detailed protocols for quantitative PCR are provided in *PCR Protocols, A Guide to Methods and Applications*, Innis *et al.*, Academic Press, Inc. N.Y., (1990).

Other suitable amplification methods include, but are not limited to polymerase chain reaction (PCR) (Innis, *et al., PCR Protocols. A guide to Methods and Application*. Academic Press, Inc. San Diego, (1990)), ligase chain reaction (LCR) (see Wu and Wallace, *Genomics*, 4: 560 (1989), Landegren, *et al., Science*, 241: 1077 (1988) and Barringer, *et al., Gene*, 89: 117 (1990), transcription amplification (Kwoh, et al., *Proc. Natl. Acad. Sci. USA*, 86: 1173 (1989)), and self-sustained sequence replication (Guatelli, *et al., Proc. Nat. Acad. Sci. USA*, 87: 1874 (1990)).

In a particularly preferred embodiment, the sample mRNA is reverse transcribed with a reverse transcriptase and a primer consisting of oligo dT and a sequence encoding the phage T7 promoter to provide single stranded DNA template. The second DNA strand is polymerized using a DNA polymerase. After synthesis of double-stranded cDNA, T7 RNA polymerase is added and RNA is transcribed from the cDNA template. Successive rounds of transcription from each single cDNA template results in amplified RNA. Methods of *in vitro* polymerization are well known to those of skill in the art (*see, e.g.*, Sambrook, *supra.*) and this particular method is described in detail by Van Gelder, *et al., Proc. Natl. Acad. Sci. USA*, 87: 1663-1667 (1990) who demonstrate that *in vitro* amplification according to this method preserves the relative frequencies of the various RNA transcripts. Moreover, Eberwine *et al. Proc. Natl. Acad. Sci. USA*, 89: 3010-3014 provide a protocol that uses two rounds of amplification

amplification, as described above, using a labeled nucleotide (*e.g.* fluorescein-labeled UTP and/or CTP) incorporates a label into the transcribed nucleic acids.

Alternatively, a label may be added directly to the original nucleic acid sample (*e.g.*, mRNA, polyA mRNA, cDNA, *etc.*) or to the amplification product after the amplification is completed. Means of attaching labels to nucleic acids are well known to those of skill in the art and include, for example nick translation or end-labeling (*e.g.* with a labeled RNA) by kinasing of the nucleic acid and subsequent attachment (ligation) of a nucleic acid linker joining the sample nucleic acid to a label (*e.g.*, a fluorophore).

Detectable labels suitable for use in the present invention include any composition detectable by spectroscopic, photochemical, biochemical, immunochemical, electrical, optical or chemical means. Useful labels in the present invention include biotin for staining with labeled streptavidin conjugate, magnetic beads (*e.g.*, Dynabeads™), fluorescent dyes (*e.g.*, fluorescein, texas red, rhodamine, green fluorescent protein, and the like), radiolabels (*e.g.*, ^{3}H, ^{125}I, ^{35}S, ^{14}C, or ^{32}P), enzymes (*e.g.*, horse radish peroxidase, alkaline phosphatase and others commonly used in an ELISA), and colorimetric labels such as colloidal gold or colored glass or plastic (*e.g.*, polystyrene, polypropylene, latex, etc.) beads. Patents teaching the use of such labels include U.S. Patent Nos. 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149; and 4,366,241.

Means of detecting such labels are well known to those of skill in the art. Thus, for example, radiolabels may be detected using photographic film or scintillation counters, fluorescent markers may be detected using a photodetector to detect emitted light. Enzymatic labels are typically detected by providing the enzyme with a substrate and detecting the reaction product produced by the action of the enzyme on the substrate, and colorimetric labels are detected by simply visualizing the colored label.

The label may be added to the target (sample) nucleic acid(s) prior to, or after the hybridization. So called "direct labels" are detectable labels that are directly attached to or incorporated into the target (sample) nucleic acid prior to hybridization. In contrast, so called "indirect labels" are joined to the hybrid duplex after hybridization.

column. After digestion with the RNase A, the hybridized DNA is removed simply by denaturing (*e.g.*, by adding heat or increasing salt) the hybrid duplexes and washing the previously hybridized mRNA off in an elution buffer.

The undigested mRNA fragments which will be hybridized to the probes in the high density array are then preferably end-labeled with a fluorophore attached to an RNA linker using an RNA ligase. This procedure produces a labeled sample RNA pool in which the nucleic acids that do not correspond to probes in the array are eliminated and thus unavailable to contribute to a background signal.

Another method of reducing sample complexity involves hybridizing the mRNA with deoxyoligonucleotides that hybridize to regions that border on either size the regions to which the high density array probes are directed. Treatment with RNAse H selectively digests the double stranded (hybrid duplexes) leaving a pool of single-stranded mRNA corresponding to the short regions (*e.g.*, 20 mer) that were formerly bounded by the deoxyoligonucleotide probes and which correspond to the targets of the high density array probes and longer mRNA sequences that correspond to regions between the targets of the probes of the high density array. The short RNA fragments are then separated from the long fragments (*e.g.*, by electrophoresis), labeled if necessary as described above, and then are ready for hybridization with the high density probe array.

In a third approach, sample complexity reduction involves the selective removal of particular (preselected) mRNA messages. In particular, highly expressed mRNA messages that are not specifically probed by the probes in the high density array are preferably removed. This approach involves hybridizing the polyA$^+$ mRNA with an oligonucleotide probe that specifically hybridizes to the preselected message close to the 3' (poly A) end. The probe may be selected to provide high specificity and low cross reactivity. Treatment of the hybridized message/probe complex with RNase H digests the double stranded region effectively removing the polyA$^+$ tail from the rest of the message. The sample is then treated with methods that specifically retain or amplify polyA$^+$ RNA (*e.g.*, an oligo dT column or (dT)n magnetic beads). Such methods will not retain or amplify the selected message(s) as they are no longer associated with a

from all other probes in the array are divided by the signal (*e.g.*, fluorescence intensity) from the control probes thereby normalizing the measurements.

Virtually any probe may serve as a normalization control. However, it is recognized that hybridization efficiency varies with base composition and probe length. Preferred normalization probes are selected to reflect the average length of the other probes present in the array, however, they can be selected to cover a range of lengths. The normalization control(s) can also be selected to reflect the (average) base composition of the other probes in the array, however in a preferred embodiment, only one or a few normalization probes are used and they are selected such that they hybridize well (*i.e.* no secondary structure) and do not match any target-specific probes.

Normalization probes can be localized at any position in the array or at multiple positions throughout the array to control for spatial variation in hybridization efficiently. In a preferred embodiment, the normalization controls are located at the corners or edges of the array as well as in the middle.

3) Expression level controls.

Expression level controls are probes that hybridize specifically with constitutively expressed genes in the biological sample. Expression level controls are designed to control for the overall health and metabolic activity of a cell. Examination of the covariance of an expression level control with the expression level of the target nucleic acid indicates whether measured changes or variations in expression level of a gene is due to changes in transcription rate of that gene or to general variations in health of the cell. Thus, for example, when a cell is in poor health or lacking a critical metabolite the expression levels of both an active target gene and a constitutively expressed gene are expected to decrease. The converse is also true. Thus where the expression levels of both an expression level control and the target gene appear to both decrease or to both increase, the change may be attributed to changes in the metabolic activity of the cell as a whole, not to differential expression of the target gene in question. Conversely, where the expression levels of the target gene and the expression level control do not covary, the variation in the expression level of the target gene is

5) Sample preparation/amplification controls.

The high density array may also include sample preparation/amplification control probes. These are probes that are complementary to subsequences of control genes selected because they do not normally occur in the nucleic acids of the particular biological sample being assayed. Suitable sample preparation/amplification control probes include, for example, probes to bacterial genes (*e.g.*, Bio B) where the sample in question is a biological from a eukaryote.

The RNA sample is then spiked with a known amount of the nucleic acid to which the sample preparation/amplification control probe is directed before processing. Quantification of the hybridization of the sample preparation/amplification control probe then provides a measure of alteration in the abundance of the nucleic acids caused by processing steps (*e.g.* PCR, reverse transcription, *in vitro* transcription, *etc.*).

B) Probe Selection and Optimization.

In a preferred embodiment, oligonucleotide probes in the high density array are selected to bind specifically to the nucleic acid target to which they are directed with minimal non-specific binding or cross-hybridization under the particular hybridization conditions utilized. Because the high density arrays of this invention can contain in excess of 1,000,000 different probes, it is possible to provide every probe of a characteristic length that binds to a particular nucleic acid sequence. Thus, for example, the high density array can contain every possible 20 mer sequence complementary to an IL-2 mRNA.

There, however, may exist 20 mer subsequences that are not unique to the IL-2 mRNA. Probes directed to these subsequences are expected to cross hybridize with occurrences of their complementary sequence in other regions of the sample genome. Similarly, other probes simply may not hybridize effectively under the hybridization conditions (*e.g.*, due to secondary structure, or interactions with the substrate or other probes). Thus, in a preferred embodiment, the probes that show such poor specificity or hybridization efficiency are identified and may not be included either in the high density array itself (*e.g.*, during fabrication of the array) or in the post-hybridization data analysis.

The high density array may additionally contain mismatch controls for each of the probes to be tested. In a preferred embodiment, the mismatch controls contain a central mismatch. Where both the mismatch control and the target probe show high levels of hybridization (*e.g.*, the hybridization to the mismatch is nearly equal to or greater than the hybridization to the corresponding test probe), the test probe is preferably not used in the high density array.

In a particularly preferred embodiment, optimal probes are selected according to the following method: First, as indicated above, an array is provided containing a multiplicity of oligonucleotide probes complementary to subsequences of the target nucleic acid. The oligonucleotide probes may be of a single length or may span a variety of lengths ranging from 5 to 50 nucleotides. The high density array may contain every probe of a particular length that is complementary to a particular mRNA or may contain probes selected from various regions of particular mRNAs. For each target-specific probe the array also contains a mismatch control probe; preferably a central mismatch control probe.

The oligonucleotide array is hybridized to a sample containing target nucleic acids having subsequences complementary to the oligonucleotide probes and the difference in hybridization intensity between each probe and its mismatch control is determined. Only those probes where the difference between the probe and its mismatch control exceeds a threshold hybridization intensity (*e.g.* preferably greater than 10% of the background signal intensity, more preferably greater than 20% of the background signal intensity and most preferably greater than 50% of the background signal intensity) are selected. Thus, only probes that show a strong signal compared to their mismatch control are selected.

The probe optimization procedure can optionally include a second round of selection. In this selection, the oligonucleotide probe array is hybridized with a nucleic acid sample that is not expected to contain sequences complementary to the probes. Thus, for example, where the probes are complementary to the RNA sense strand a sample of antisense RNA is provided. Of course, other samples could be provided such as samples from organisms or cell lines known to be lacking a particular gene, or known for not expressing a particular gene.

6) Clumping score is less than 6.

7) Number of As + Number of Ts is less than 14

8) Number of As + number of Gs is less than 15

With respect to rule number 4, requiring the maximum run of any two bases to be less than 11 bases guarantees that at least three different bases occur within any 12 consecutive nucleotides. A palindrome score is the maximum number of complementary bases if the oligonucleotide is folded over at a point that maximizes self complementarity. Thus, for example a 20 mer that is perfectly self-complementary would have a palindrome score of 10. A clumping score is the maximum number of three-mers of identical bases in a given sequence. Thus, for example, a run of 5 identical bases will produce a clumping score of 3 (bases 1-3, bases 2-4, and bases 3-5).

If any probe failed one of these criteria (1-8), the probe was not a member of the subset of probes placed on the chip. For example, if a hypothetical probe was 5'-AGCTTTTTTCATGCATCTAT-3' the probe would not be synthesized on the chip because it has a run of four or more bases (*i.e.*, run of six).

The cross hybridization rules developed for 20 mers were as follows:

1) Number of Cs is less than 8;

2) Number of Cs in any window of 8 bases is less than 4.

Thus, if any probe failed any of either the hybridization ruses (1-8) or the cross-hybridization rules (1-2), the probe was not a member of the subset of probes placed on the chip. These rules eliminated many of the probes that cross hybridized strongly or exhibited low hybridization, and performed moderate job of eliminating weakly hybridizing probes.

These heuristic rules may be implemented by hand calculations, or alternatively, they may be implemented in software as is discussed below in Section IV.B.7.

3) Neural net.

In another embodiment, a neural net can be trained to predict the hybridization and cross-hybridization intensities based on the sequence of the probe or on other probe properties. The neural net can then be used to pick an arbitrary number

```
probe from gene 1:  aagcgcgatcgattatgctc
                    |      |||||||
gene 2:             atctcggatcgatcggataagcgcgatcgattatgctcggcga
```

has 8 matching bases in this alignment, but 20 matching bases in the following alignment:

```
probe from gene 1:                    aagcgcgatcgattatgctc
                                      ||||||||||||||||||||
gene 2:             atctcggatcgatcggataagcgcgatcgattatgctcggcga
```

More complicated algorithms also exist, which allow the detection of insertion or deletion mismatches. Such sequence alignment algorithms are well known to those of skill in the art and include, but are not limited to BLAST, or FASTA, or other gene matching programs such as those described above in the definitions section.

In another variant, where an organism has many different genes which are very similar, it is difficult to make a probe set that measures the concentration only one of those very similar genes. One can then prune out any probes which are dissimilar, and make the probe set a probe set for that family of genes.

6) Synthesis cycle pruning.

The cost of producing masks for a chip is approximately linearly related to the number of synthesis cycles. In a normal set of genes the distribution of the number of cycles any probe takes to build approximates a Gausian distribution. Because of this the mask cost can normally be reduced by 15% by throwing out about 3 percent of the probes. In a preferred embodiment, synthesis cycle pruning simply involves eliminating (not including) those probes those probes that require a greater number of synthesis cycles than the maximum number of synthesis cycles selected for preparation of the particular subject high density oligonucleotide array. Since the typical synthesis of probes follows a regular pattern of bases put down (acgtacgtacgt...) counting the number of synthesis steps needed to build a probe is easy. The listing shown in Table 1 povides typical code for counting the number of synthesis cycles a probe will need.

do not necessarily produce the same results, or produce entirely independent results, it may be advantageous to combine the methods. For example, probes may be pruned or reduced if more than one method (*e.g.*, two out of three) indicate the probe will not likely produce good results. Then, synthesis cycle pruning may be performed to reduce costs.

Fig. 11 shows the flow of a process of increasing the number of probes for monitoring the expression of genes after the number of probes has been reduced or pruned. In one embodiment, a user is able to specify the number of nucleic acid probes that should be placed on the chip to monitor the expression of each gene. As discussed above, it is advantageous to reduce probes that will not likely produce good results; however, the number of probes may be reduced to substantially less than the desired number of probes.

At step 402, the number of probes for monitoring multiple genes is reduced by the heuristic rules method, neural net, annova model, synthesis cycle pruning, or any other method, or combination of methods. A gene is selected at step 404.

A determination is made whether the remaining probes for monitoring the selected gene number greater than 80% (which may be varied or user defined) of the desired number of probes. If yes, the computer system proceeds to the next gene at step 408 which will generally return to step 404.

If the remaining probes for monitoring the selected gene do not number greater than 80% of the desired number of probes, a determination is made whether the remaining probes for monitoring the selected gene number greater than 40% (which may be varied or user defined) of the desired number of probes. If yes, an "i" is appended to the end of the gene name to indicate that after pruning, the probes were incomplete at step 412.

At step 414, the number of probes is increased by loosening the constraints that rejected probes. For example, the thresholds in the heuristic rules may be increased by 1. Therefore, if previously probes were rejected if they had four As in a row, the rule may be loosened to five As in a row.

A determination is then made whether the remaining probes for monitoring the selected gene number greater than 80% of the desired number of probes at step 416. If yes, an "r" is appended to the end of the gene name at step 412 to indicate that the rules were loosened to generate the number of synthesized probes for that gene.

simultaneous coupling at a number of reaction sites, into a different heterogenous array. See, U.S. Application Serial Nos. 07/796,243 and 07/980,523.

The development of VLSIPS™ technology as described in the above-noted U.S. Patent No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092, is considered pioneering technology in the fields of combinatorial synthesis and screening of combinatorial libraries. More recently, patent application Serial No. 08/082,937, filed June 25, 1993 describes methods for making arrays of oligonucleotide probes that can be used to check or determine a partial or complete sequence of a target nucleic acid and to detect the presence of a nucleic acid containing a specific oligonucleotide sequence.

In brief, the light-directed combinatorial synthesis of oligonucleotide arrays on a glass surface proceeds using automated phosphoramidite chemistry and chip masking techniques. In one specific implementation, a glass surface is derivatized with a silane reagent containing a functional group, *e.g.*, a hydroxyl or amine group blocked by a photolabile protecting group. Photolysis through a photolithogaphic mask is used selectively to expose functional groups which are then ready to react with incoming 5'-photoprotected nucleoside phosphoramidites. The phosphoramidites react only with those sites which are illuminated (and thus exposed by removal of the photolabile blocking group). Thus, the phosphoramidites only add to those areas selectively exposed from the preceding step. These steps are repeated until the desired array of sequences have been synthesized on the solid surface. Combinatorial synthesis of different oligonucleotide analogues at different locations on the array is determined by the pattern of illumination during synthesis and the order of addition of coupling reagents.

In the event that an oligonucleotide analogue with a polyamide backbone is used in the VLSIPS™ procedure, it is generally inappropriate to use phosphoramidite chemistry to perform the synthetic steps, since the monomers do not attach to one another via a phosphate linkage. Instead, peptide synthetic methods are substituted. See, *e.g.*, Pirrung *et al.* U.S. Pat. No. 5,143,854.

Peptide nucleic acids are commercially available from, *e.g.*, Biosearch, Inc. (Bedford, MA) which comprise a polyamide backbone and the bases found in naturally occurring nucleosides. Peptide nucleic acids are capable of binding to nucleic

the monomer B is flowed through or placed in the second flow channel(s), binding monomer B at the second selected locations. In this particular example, the resulting sequences bound to the substrate at this stage of processing will be, for example, A, B, and AB. The process is repeated to form a vast array of sequences of desired length at known locations on the substrate.

After the substrate is activated, monomer A can be flowed through some of the channels, monomer B can be flowed through other channels, a monomer C can be flowed through still other channels, *etc.* In this manner, many or all of the reaction regions are reacted with a monomer before the channel block must be moved or the substrate must be washed and/or reactivated. By making use of many or all of the available reaction regions simultaneously, the number of washing and activation steps can be minimized.

One of skill in the art will recognize that there are alternative methods of forming channels or otherwise protecting a portion of the surface of the substrate. For example, according to some embodiments, a protective coating such as a hydrophilic or hydrophobic coating (depending upon the nature of the solvent) is utilized over portions of the substrate to be protected, sometimes in combination with materials that facilitate wetting by the reactant solution in other regions. In this manner, the flowing solutions are further prevented from passing outside of their designated flow paths.

The "spotting" methods of preparing compounds and libraries of the present invention can be implemented in much the same manner as the flow channel methods. For example, a monomer A can be delivered to and coupled with a first group of reaction regions which have been appropriately activated. Thereafter, a monomer B can be delivered to and reacted with a second group of activated reaction regions. Unlike the flow channel embodiments described above, reactants are delivered by directly depositing (rather than flowing) relatively small quantities of them in selected regions. In some steps, of course, the entire substrate surface can be sprayed or otherwise coated with a solution. In preferred embodiments, a dispenser moves from region to region, depositing only as much monomer as necessary at each stop. Typical dispensers include a micropipette to deliver the monomer solution to the substrate and a robotic system to control the position of the micropipette with respect to the substrate.

signal intensity greater than approximately 10% of the background intensity. Thus, in a preferred embodiment, the hybridized array may be washed at successively higher stringency solutions and read between each wash. Analysis of the data sets thus produced will reveal a wash stringency above which the hybridization pattern is not appreciably altered and which provides adequate signal for the particular oligonucleotide probes of interest.

In a preferred embodiment, background signal is reduced by the use of a detergent (*e.g.*, C-TAB) or a blocking reagent (*e.g.*, sperm DNA, cot-1 DNA, etc.) during the hybridization to reduce non-specific binding. In a particularly preferred embodiment, the hybridization is performed in the presence of about 0.5 mg/ml DNA (*e.g.*, herring sperm DNA). The use of blocking agents in hybridization is well known to those of skill in the art (*see, e.g.*, Chapter 8 in P. Tijssen, *supra.*)

The stability of duplexes formed between RNAs or DNAs are generally in the order of RNA:RNA > RNA:DNA > DNA:DNA, in solution. Long probes have better duplex stability with a target, but poorer mismatch discrimination than shorter probes (mismatch discrimination refers to the measured hybridization signal ratio between a perfect match probe and a single base mismatch probe). Shorter probes (*e.g.*, 8-mers) discriminate mismatches very well, but the overall duplex stability is low.

Altering the thermal stability (T_m) of the duplex formed between the target and the probe using, *e.g.*, known oligonucleotide analogues allows for optimization of duplex stability and mismatch discrimination. One useful aspect of altering the T_m arises from the fact that adenine-thymine (A-T) duplexes have a lower T_m than guanine-cytosine (G-C) duplexes, due in part to the fact that the A-T duplexes have 2 hydrogen bonds per base-pair, while the G-C duplexes have 3 hydrogen bonds per base pair. In heterogeneous oligonucleotide arrays in which there is a non-uniform distribution of bases, it is not generally possible to optimize hybridization for each oligonucleotide probe simultaneously. Thus, in some embodiments, it is desirable to selectively destabilize G-C duplexes and/or to increase the stability of A-T duplexes. This can be accomplished, *e.g.*, by substituting guanine residues in the probes of an array which form G-C duplexes with hypoxanthine, or by substituting adenine residues in probes

embodiment, the excitation light source is a laser appropriate for the excitation of the fluorescent label.

The confocal microscope may be automated with a computer-controlled stage to automatically scan the entire high density array. Similarly, the microscope may be equipped with a phototransducer (*e.g.*, a photomultiplier, a solid state array, a ccd camera, *etc.*) attached to an automated data acquisition system to automatically record the fluorescence signal produced by hybridization to each oligonucleotide probe on the array. Such automated systems are described at length in U.S. Patent No: 5,143,854, PCT Application 20 92/10092, and copending U.S.S.N. 08/195,889 filed on February 10, 1994. Use of laser illumination in conjunction with automated confocal microscopy for signal detection permits detection at a resolution of better than about 100 μm, more preferably better than about 50 μm, and most preferably better than about 25 μm.

VIII. Signal Evaluation.

One of skill in the art will appreciate that methods for evaluating the hybridization results vary with the nature of the specific probe nucleic acids used as well as the controls provided. In the simplest embodiment, simple quantification of the fluorescence intensity for each probe is determined. This is accomplished simply by measuring probe signal strength at each location (representing a different probe) on the high density array (*e.g.*, where the label is a fluorescent label, detection of the amount of florescence (intensity) produced by a fixed excitation illumination at each location on the array). Comparison of the absolute intensities of an array hybridized to nucleic acids from a "test" sample with intensities produced by a "control" sample provides a measure of the relative expression of the nucleic acids that hybridize to each of the probes.

One of skill in the art, however, will appreciate that hybridization signals will vary in strength with efficiency of hybridization, the amount of label on the sample nucleic acid and the amount of the particular nucleic acid in the sample. Typically nucleic acids present at very low levels (*e.g.*, < 1pM) will show a very weak signal. At some low level of concentration, the signal becomes virtually indistinguishable from background. In evaluating the hybridization data, a threshold intensity value may be

corresponding test probe, there is a problem with the hybridization and the signal from those probes is ignored. The difference in hybridization signal intensity between the target specific probe and its corresponding mismatch control is a measure of the discrimination of the target-specific probe. Thus, in a preferred embodiment, the signal of the mismatch probe is subtracted from the signal from its corresponding test probe to provide a measure of the signal due to specific binding of the test probe.

The concentration of a particular sequence can then be determined by measuring the signal intensity of each of the probes that bind specifically to that gene and normalizing to the normalization controls. Where the signal from the probes is greater than the mismatch, the mismatch is subtracted. Where the mismatch intensity is equal to or greater than its corresponding test probe, the signal is ignored. The expression level of a particular gene can then be scored by the number of positive signals (either absolute or above a threshold value), the intensity of the positive signals (either absolute or above a selected threshold value), or a combination of both metrics (*e.g.*, a weighted average).

It is a surprising discovery of this invention, that normalization controls are often unnecessary for useful quantification of a hybridization signal. Thus, where optimal probes have been identified in the two step selection process as described above, in Section II.B., the average hybridization signal produced by the selected optimal probes provides a good quantified measure of the concentration of hybridized nucleic acid.

IX. Computer-implemented Expression Monitoring

The methods of monitoring gene expression of this invention may be performed utilizing a computer. The computer typically runs a software program that includes computer code incorporating the invention for analyzing hybridization intensities measured from a substrate or chip and thus, monitoring the expression of one or more genes. Although the following will describe specific embodiments of the invention, the invention is not limited to any one embodiment so the following is for purposes of illustration and not limitation.

chip. Most preferably, the nucleic acid probes have a density greater than about 60 different nucleic acid probes per 1 cm^2 of the substrate. Although the flowcharts show a sequence of steps for clarity, this is not an indication that the steps must be performed in this specific order. One of ordinary skill in the art would readily recognize that many of the steps may be reordered, combined, and deleted without departing from the invention.

Initially, nucleic acid probes are selected that are complementary to the target sequence (or gene). These probes are the perfect match probes. Another set of probes is specified that are intended to be not perfectly complementary to the target sequence. These probes are the mismatch probes and each mismatch probe includes at least one nucleotide mismatch from a perfect match probe. Accordingly, a mismatch probe and the perfect match probe from which it was derived make up a pair of probes. As mentioned earlier, the nucleotide mismatch is preferably near the center of the mismatch probe.

The probe lengths of the perfect match probes are typically chosen to exhibit high hybridization affinity with the target sequence. For example, the nucleic acid probes may be all 20-mers. However, probes of varying lengths may also be synthesized on the substrate for any number of reasons including resolving ambiguities.

The target sequence is typically fragmented, labeled and exposed to a substrate including the nucleic acid probes as described earlier. The hybridization intensities of the nucleic acid probes is then measured and input into a computer system. The computer system may be the same system that directs the substrate hybridization or it may be a different system altogether. Of course, any computer system for use with the invention should have available other details of the experiment including possibly the gene name, gene sequence, probe sequences, probe locations on the substrate, and the like.

Referring to Fig. 8, after hybridization, the computer system receives input of hybridization intensities of the multiple pairs of perfect match and mismatch probes at step 202. The hybridization intensities indicate hybridization affinity between the nucleic acid probes and the target nucleic acid (which corresponds to a gene). Each pair includes a perfect match probe that is perfectly complementary to a portion of the

are typically user defined values that have been determined to produce accurate expression monitoring of a gene or genes. In one embodiment, the difference threshold is 20 and the ratio threshold is 1.2.

If $I_{pm} - I_{mm} >= D$ and $I_{pm} / I_{mm} >= R$, the value NPOS is incremented at step 260. In general, NPOS is a value that indicates the number of pairs of probes which have hybridization intensities indicating that the gene is likely expressed. NPOS is utilized in a determination of the expression of the gene.

At step 262, it is determined if $I_{mm} - I_{pm} >= D$ and $I_{mm} / I_{pm} >= R$. If this expression is true, the value NNEG is incremented at step 264. In general, NNEG is a value that indicates the number of pairs of probes which have hybridization intensities indicating that the gene is likely not expressed. NNEG, like NPOS, is utilized in a determination of the expression of the gene.

For each pair that exhibits hybridization intensities either indicating the gene is expressed or not expressed, a log ratio value (LR) and intensity difference value (IDIF) are calculated at step 266. LR is calculated by the log of the quotient of the hybridization intensities of the pair (I_{pm} / I_{mm}). The IDIF is calculated by the difference between the hybridization intensities of the pair ($I_{pm} - I_{mm}$). If there is a next pair of hybridization intensities at step 268, they are retrieved at step 254.

At step 272, a decision matrix is utilized to indicate if the gene is expressed. The decision matrix utilizes the values N, NPOS, NNEG, and LR (multiple LRs). The following four assignments are performed:

P1 = NPOS / NNEG

P2 = NPOS / N

P3 = (10 * SUM(LR)) / (NPOS + NNEG)

These P values are then utilized to determine if the gene is expressed.

For purposes of illustration, the P values are broken down into ranges. If P1 is greater than or equal to 2.1, then A is true. If P1 is less than 2.1 and greater than or equal to 1.8, then B is true. Otherwise, C is true. Thus, P1 is broken down into three ranges A, B and C. This is done to aid the readers understanding of the invention.

Thus, all of the P values are broken down into ranges according to the following:

Once all the pairs of probes have been processed and the expression of the gene indicated, an average of ten times the LRs is computed at step 275. Additionally, an average of the IDIF values for the probes that incremented NPOS and NNEG is calculated. These values may be utilized for quantitative comparisons of this experiments with other experiments.

Quantitative measurements may be performed at step 276. For example, the current experiment may be compared to a previous experiment (*e.g.*, utilizing values calculated at step 270). Additionally, the experiment may be compared to hybridization intensities of RNA (such as from bacteria) present in the biological sample in a known quantity. In this manner, one may verify the correctness of the gene expression indication or call, modify threshold values, or perform any number of modifications of the preceding.

For simplicity, Fig. 9 was described in reference to a single gene. However, the process may be utilized on multiple genes in a biological sample. Therefore, any discussion of the analysis of a single gene is not an indication that the process may not be extended to processing multiple genes.

Figs. 10A and 10B show the flow of a process of determining the expression of a gene by comparing baseline scan data and experimental scan data. For example, the baseline scan data may be from a biological sample where it is known the gene is expressed. Thus, this scan data may be compared to a different biological sample to determine if the gene is expressed. Additionally, it may be determined how the expression of a gene or genes changes over time in a biological organism.

At step 302, the computer system receives raw scan data of N pairs of perfect match and mismatch probes from the baseline. The hybridization intensity of a perfect match probe from the baseline will be designed "I_{pm}" and the hybridization intensity of a mismatch probe from the baseline will be designed "I_{mm}." The background signal intensity is subtracted from each of the hybridization intensities of the pairs of baseline scan data at step 304.

At step 306, the computer system receives raw scan data of N pairs of perfect match and mismatch probes from the experimental biological sample. The hybridization intensity of a perfect match probes from the experiment will be designed

been added to each value in order to indicate if the value denotes the baseline sample or the experimental sample, respectively. If there are next pairs of hybridization intensities at step 322, they are processed in a similar manner as shown.

Referring now to Fig. 10B, an absolute decision computation is performed for both the baseline and experimental samples at step 324. The absolute decision computation is an indication of whether the gene is expressed, marginal or absent in each of the baseline and experimental samples. Accordingly, in a preferred embodiment, this step entails performing steps 272 and 274 from Fig. 9 for each of the samples. This being done, there is an indication of gene expression for each of the samples taken alone.

At step 326, a decision matrix is utilized to determine the difference in gene expression between the two samples. This decision matrix utilizes the values, N, NPOSB, NPOSE, NNEGB, NNEGE, NINC, NDEC, LRB, and LRE as they were calculated above. The decision matrix performs different calculations depending on whether NINC is greater than or equal to NDEC. The calculations are as follows.

If NINC > = NDEC, the following four P values are determined:

P1 = NINC / NDEC

P2 = NINC / N

P3 = ((NPOSE - NPOSB) - (NNEGE - NNEGB)) / N

P4 = 10 * SUM(LRE - LRB) / N

These P values are then utilized to determine the difference in gene expression between the two samples.

For purposes of illustration, the P values are broken down into ranges as was done previously. Thus, all of the P values are broken down into ranges according to the following:

A = (P1 > = 2.7)

B = (2.7 > P1 > = 1.8)

C = (P1 < 1.8)

X = (P2 > = 0.24)

P1 = NDEC / NINC

P2 = NDEC / N

P3 = ((NNEGE - NNEGB) - (NPOSE - NPOSB)) / N

P4 = 10 * SUM(LRE - LRB) / N

These P values are then utilized to determine the difference in gene expression between the two samples.

The P values are broken down into the same ranges as for the other case where NINC > = NDEC. Thus, P values in this case indicate the same ranges and will not be repeated for the sake of brevity. However, the ranges generally indicate different changes in the gene expression between the two samples as shown below.

In this case where NINC < NDEC, the gene expression change is indicated as decreased, marginal decrease or no change. The following is a summary of the gene expression indications:

Decreased	A and (X or Y) and (Q or R) and (M or N or O)
	A and (X or Y) and (Q or R or S) and (M or N)
	B and (X or Y) and (Q or R) and (M or N)
	A and X and (Q or R or S) and (M or N or O)
Marginal	A or Y or S or O
Decrease	B and (X or Y) and (Q or R) and O
	B and (X or Y) and S and (M or N)
	C and (X or Y) and (Q or R) and (M or N)
No Change	All others cases (*e.g.*, any Z combination)

In the output to the user, decreased may be indicated as "D," marginal decrease as "MD" and no change as "NC."

The above has shown that the relative difference between the gene expression between a baseline sample and an experimental sample may be determined.

levels of the marker in the test sample as compared to the healthy sample. Correlation of such deviations with a pathological condition provides a diagnostic assay for that condition.

EXAMPLES

The following examples are offered to illustrate, but not to limit the present invention.

Example 1

First Generation Oligonucleotide Arrays Designed to Measure mRNA Levels for a Small Number of Murine Cytokines.

A) Preparation of labeled RNA.

1) From each of the preselected genes.

Fourteen genes (IL-2, IL-3, Il-4, IL-6, Il-10, IL-12p40, GM-CSF, IFN-γ, TNF-α, CTLA8, ß-actin, GAPDH, IL-11 receptor, and Bio B) were each cloned into the p Bluescript II KS (+) phagemid (Stratagene, La Jolla, California, USA). The orientation of the insert was such that T3 RNA polymerase gave sense transcripts and T7 polymerase gave antisense RNA.

Labeled ribonucleotides in an in vitro transcription (IVT) reaction. Either biotin- or fluorescein-labeled UTP and CTP (1:3 labeled to unlabeled) plus unlabeled ATP and GTP were used for the reaction with 2500 units of T7 RNA polymerase (Epicentre Technologies, Madison, Wisconsin, USA). *In vitro* transcription was done with cut templates in a manner like that described by Melton *et al.*, *Nucleic Acids Research*, 12: 7035-7056 (1984). A typical *in vitro* transcription reaction used 5 μg DNA template, a buffer such as that included in Ambion's Maxiscript *in vitro* Transcription Kit (Ambion Inc., Huston, Texas, USA) and GTP (3 mM), ATP (1.5 mM), and CTP and fluoresceinated UTP (3 mM total, UTP: Fl-UTP 3:1) or UTP and fluoresceinated CTP (2 mM total, CTP: Fl-CTP, 3:1). Reactions done in the Ambion buffer had 20 mM DTT and RNase inhibitor. The reaction was run from 1.5 to about 8 hours.

Following the reaction, unincorporated nucleotide triphosphates were removed using a size-selective membrane (microcon-100) or Pharmacia microspin S-200

IL-10 at 1:35,000. Labeled sense and antisense T10 RNA samples were synthesized from NotI and SfiI cut CDNA libraries in *in vitro* transcription reactions as described above.

ii) The 2D6 murine helper T cells line.

The 2D6 cell line is a murine IL-12 dependent T cell line developed by Fujiwara *et al.* Cells were cultured in RPMI 1640 medium with 10% heat inactivated fetal calf serum (JRH Biosciences), 0.05 mM P-mercaptoethanol and recombinant murine IL-12 (100 units/mL, Genetics Institute, Cambridge, Massachusetts, USA). For cytokine induction, cells were preincubated overnight in IL-12 free medium and then resuspended (10^6 cells/ml). After incubation for 0, 2, 6 and 24 hours in media containing 5 nM calcium ionophore A23187 (Sigma Chemical Co., St. Louis Missouri, USA) and 100 nM 4-phorbol-12-myristate 13-acetate (Sigma), cells were collected by centrifugation and washed once with phosphate buffered saline prior to isolation of RNA.

Labeled 2D6 mRNA was produced by directionally cloning the 2D6 cDNA with αZipLox, NotI-SalI arms available from GibcoBRL in a manner similar to T10. The linearized pZl1 library was transcribed with T7 to generate sense RNA as described above.

iii) RNA preparation.

For material made directly from cellular RNA, cytoplasmic RNA was extracted from cells by the method of Favaloro *et al.*, (1980) *Meth. Enzym.*, 65: 718-749, and poly $(A)^+$ RNA was isolated with an oligo dT selection step (PolyAtract, Promega,). RNA was amplified using a modification of the procedure described by Eberwine *et al.* (1992) *Proc. Natl. Acad. Sci. USA, 89: 3010-3014* (*see also* Van Gelder *et al.* (1990) *Science* 87: 1663-1667). One microgram of poly (A)+ RNA was converted into double-stranded cDNA using a cDNA synthesis kit (Life Technologies) with an oligo dT prime incorporating a T7 RNA polymerase promoter site. After second strand synthesis, the reaction mixture was extracted with phenol/chloroform and the double-stranded DNA isolated using a membrane filtration step (Mircocon-100,

Table 2. High density array design. For every probe there was also a mismatch control having a central 1 base mismatch.

Probe Type	Target Nucleic Acid	Number of Probes
Test Probes:	IL-2	691
	IL-3	751
	IL-4	361
	IL-6	691
	IL-10	481
	IL-12p40	911
	GM-CSF	661
	IFN-γ	991
	TNF-α	641
	mCTLA8	391
	IL-11 receptor	158
House Keeping Genes:	GAPDH	388
	ß-actin	669
Bacterial gene (sample preparation/amplification control)	Bio B	286

The high density array was synthesized on a planar glass slide.

C) Array hybridization and scanning.

The RNA transcribed from cDNA was hybridized to the high density oligonucleotide probe array(s) at low stringency and then washed under more stringent conditions. The hybridization solutions contained 0.9 M NaCl, 60 mM NaH_2PO_4, 6 mM EDTA and 0.005 % Triton X-100 , adjusted to pH 7.6 (referred to as 6x SSPE-T). In addition, the solutions contained 0.5 mg/ml unlabeled, degraded herring sperm DNA (Sigma Chemical Co., St. Louis, Missouri, USA). Prior to hybridization, RNA samples were heated in the hybridization solution to 9 "C for 10 minutes, placed on ice for 5 minutes, and allowed to equilibrate at room temperature before being placed in the hybridization flow cell, Following hybridization, the solution was removed, the arrays

D) Optimization of Probe Selection

In order to optimize probe selection for each of the target genes, the high density array of oligonucleotide probes was hybridized with the mixture of labeled RNAs transcribed from each of the target genes. Fluorescence intensity at each location on the high density array was determined by scanning the high density array with a laser illuminated scanning confocal fluorescence microscope connected to a data acquisition system.

Probes were then selected for further data analysis in a two-step procedure. First, in order to be counted, the difference in intensity between a probe and its corresponding mismatch probe had to exceed a threshold limit (50 counts, or about half background, in this case). This eliminated from consideration probes that did not hybridize well and probes for which the mismatch control hybridizes at an intensity comparable to the perfect match.

The high density array was hybridized to a labeled RNA sample which, in principle, contains none of the sequences on the high density array. In this case, the oligonucleotide probes were chosen to be complementary to the sense RNA. Thus, an anti-sense RNA population should have been incapable of hybridizing to any of the probes on the array. Where either a probe or its mismatch showed a signal above a threshold value (100 counts above background) it was not included in subsequent analysis.

Then, the signal for a particular gene was counted as the average difference (perfect match - mismatch control) for the selected probes for each gene.

E) Results: The high density arrays provide specific and sensitive detection of target nucleic acids.

As explained above, the initial arrays contained more than 16,000 probes that were complementary to 12 murine mRNAs - 9 cytokines, 1 cytokine receptor, 2 constitutively expressed genes (5-actin and glyceraldehyde 3-phosphate dehydrogenase) - 1 rat cytokine and 1 bacterial gene *(E. coli* biotin synthetase, *bioB*) which serves as a quantitation reference. The initial experiments with these relatively simple arrays were designed to determine whether short *in situ* synthesized oligonucleotides can be made to

mCTLA8 and Bio B. The hybridized array showed strong specific signals for each of the test target nucleic acids with minimal cross hybridization.

2) Detection of Gene Expression levels in a complex target sample.

To determine how well individual RNA targets could be detected in the presence of total mammalian cell message populations, spiking experiments were carried out. Known amounts of individual RNA targets were spiked into labeled RNA derived from a representative cDNA library made from the murine B cell line T10. The T10 cell line was chosen because of the cytokines being monitored, only IL-10 is expressed at a detectable level.

Because simply spiking the RNA mixture with the selected target genes and then immediately hybridizing might provide an artificially elevated reading relative to the rest of the mixture, the spiked sample was treated to a series of procedures to mitigate differences between the library RNA and the added RNA. Thus the "spike" was added to the sample which was then heated to 37°C and annealed. The sample was then frozen, thawed, boiled for 5 minutes, cooled on ice and allowed to return to room temperature before performing the hybridization.

Figure 2A shows the results of an experiment in which 13 target RNAS were spiked into the total RNA pool at a level of 1:3000 (equivalent to a few hundred copies per cell). RNA frequencies are given as the molar amount of an individual RNA per mole of total RNA. Figure 2B shows a small portion of the array (the boxed region of 2A) containing probes specific for interleukin-2 and interleukin-3 (IL-2 and IL-3,) RNA, and Figure 2C shows the same region in the absence of the spiked targets. The hybridization signals are specific as indicated by the comparison between the spiked and unspiked images, and perfect match (PM) hybridizations are well discriminated from missmatches (MM) as shown by the pattern of alternating brighter rows (corresponding to PM probes) and darker rows (corresponding to MM probes). The observed variation among the different perfect match hybridization signals was highly reproducible and reflects the sequence dependence of the hybridizations. In a few instances, the perfect match (PM) probe was not significantly brighter than its mismatch (MM) partner because of cross-hybridization with other members of the complex RNA population.

contained 0.05 mg/ml of sense RNA transcribed from the murine B cell CDNA library, plus combinations of the 10 cytokine RNAs each at a different concentration. Individual cytokine RNAs were spiked at one of the following levels: 0, 1:300,000, 1:30,000, 1:3000, or 1:300. The four samples plus an unspiked reference were hybridized to separate arrays for 15 hours at 40°C. The presence or absence of an RNA target was determined by the pattern of hybridization and how it differed from that of the unspiked reference, and the concentrations were detected by the intensities. The concentrations of each of the ten cytokines in the four blind samples were correctly determined, with no false positives or false negatives.

One case is especially noteworthy: IL-10 is expressed in the mouse B cells used to make the CDNA library, and was known to be present in the library at a frequency of 1:60,000 to 1:30,000. In one of the unknowns, an additional amount of IL-10 RNA (corresponding to a frequency of 1:300,000) was spiked into the sample. The amount of the spiked IL-10 RNA was correctly determined, even though it represented an increase of only 10 - 20% above the intrinsic level. These results indicate that subtle changes in expression are sensitively determined by performing side-by-side experiments with identically prepared samples on identically synthesized arrays.

Example 2

T Cell Induction Experiments Measuring Cytokine mRNAs as a Function of Time Following Stimulation.

The high density arrays of this invention were next used to monitor cytokine MRNA levels in murine T cells at different times following a biochemical stimulus. Cells from the murine T helper cell line (2D6) were treated with the phorbol ester 4-phorbol-12-myristate 13-acetate (PMA) and a calcium ionophore. Poly $(A)^+$ MRNA was then isolated at 0, 2, 6 and 24 hours after stimulation. Isolated mRNA (approximately 1 μg) was converted to labeled antisense RNA using a procedure that combines a double-stranded cDNA synthesis step with a subsequent *in vitro* transcription reaction. This RNA synthesis and labeling procedure amplifies the entire mRNA

obtained for the set of genes in common to both chip types. Expression changes were unambiguously observed for more than 20 other genes in addition to those shown in Figure 4.

To determine whether much smaller sets of probes per gene are sufficient for reliable detection of RNAs, hybridization results from the 118 gene chip were analyzed using ten different subsets of 20 probe pairs per gene. That is to say, the data were analyzed as if the arrays contained only 20 probe pairs per gene. The ten subsets of 20 pairs were chosen from the approximately 300 probe pairs per gene on the arrays. The initial probe selection was made utilizing the probe selection and pruning algorithms described above. The ten subjects of 20 pairs were then randomly chosen from those probes that survived selection and pruning. Labeled RNAs were spiked into the murine B cell RNA population at levels of 1:25,000, 1:50,000 and 1:100,000. Changes in hybridization signals for the spiked RNAs were consistently detected at all three levels with the smaller probe sets. As expected, the hybridization intensities do not cluster as tightly as when averaging over larger numbers of probes. This analysis indicates that sets of 20 probe pairs per gene are sufficient for the measurement of expression changes at low levels, but that improvements in probe selection and experimental procedures will are preferred to routinely detect RNAs at the very lowest levels with such small probe sets. Such improvements include, but are not limited to higher stringency hybridizations coupled with use of slightly longer oligonucleotide probes (*e.g.*, 25 mer probes)) are in progress.

Example 4

Scale Up to Thousands of Genes

A set of four high density arrays each containing 25-mer oligonucleotide probes approximately 1650 different human genes provided probes to a total of 6620 genes. There were about 20 probes for each gene. The feature size on arrays was 50 microns. This high density array was successfully hybridized to a cDNA library using essentially the protocols described above. Similar sets of high density arrays containing oligonucleotide probes to every known expressed sequence tag (EST) are in preparation.

between predicted intensity and measured intensity, with a better model for cross hybridization than hybridization.

A) Input/output mapping.

The neural net was trained to identify the hybridization properties of 20-mer probes. The 20-mer probes were mapped to an eighty bit long input vector, with the first four bits representing the base in the first position of the probe, the next four bits representing the base in the second position, *etc*. Thus, the four bases were encoded as follows:

A: 1000

C: 0100

G: 0010

T: 0001

The neural network produced two outputs; hybridization intensity, and crosshybridization intensity. The output was scaled linearly so that 95% of the outputs from the actual experiments fell in the range 0. to 1.

B) Neural net architecture.

The neural net was a backpropagation network with 80 input neurons, one hidden layer of 20 neurons, and an output layer of two neurons. A sigmoid transfer function was used: ($s(x) = 1/(1+ \exp(-1 * x))$) that scales the input values from 0 to 1 in a non-linear (sigmoid) manner.

C) Neural net training.

The network was trained using the default parameters from Neural Works Professional 2.5 for a backprop network. (Neural Works Professional is a product of NeuralWare, Pittsburgh Pennsylvania, USA). The training set consisted of approximately 8000 examples of probes, and the associated hybridization and crosshybridization intensities.

0.33042327 0.4072904 0.24270254 -0.3750777 0.14083703 0.30998308
0.19591335 -0.4028497 0.30585453 0.35896543 0.24851802 -0.2937264
0.19672842 0.16133355 0.21780767 -0.2419563 0.17847325 0.07593013
0.1710967 -0.2728708 0.1234024 0.06987085 0.1741322 0.05922241
0.03326527 0.22045346 0.98782647=

-0.0752053 -0.0571054 -0.1834571 0.14263187 -0.0715346 -0.0524248
-0.0838031 0.01667063 -0.0945634 -0.1137057 -0.1040308 0.04263301
-0.2039919 -0.0532526 -0.0828366 0.1373803 -0.0562212 -0.2127942
-0.0482095 0.04316666 -0.1732933 0.0550463 -0.0526818 0.06739104
-0.0065265 -0.2011867 -0.0434558 -0.0369132 -0.0196296 -0.1314755
0.09420983 -0.0010159 -0.1768979 -0.2365085 -0.0150508 0.14120786
0.00565713 -0.1990354 0.11568499 -0.0690084 -0.1509431 -0.0575663
0.11275655 0.01772332 -0.0016695 -0.249011 0.09066539 0.05357879
-0.0850152 -0.1931012 0.08498721 0.03673514 -0.1446398 -0.199778
0.1065109 0.07205399 -0.1304159 -0.1723315 0.09151162 0.05596334
-0.0922655 -0.1478272 0.08858409 0.14206541 -0.0314846 -0.1985286
0.19862956 -0.0502828 -0.11447 -0.1440073 0.01366408 0.11101657
-0.0721622 -0.1506944 0.14910588 0.03297219 -0.0266356 -0.2501774
0.20344114 -0.061502 -0.1647823=

0.02848385 0.00254791 -0.0646306 0.02634032 -0.0654473 0.04731949
-0.0742345 -0.0545447 -0.1119258 0.10765317 -0.0606677 0.05693235
-0.0747124 0.13325705 -0.0508435 -0.1761459 -0.0883804 -0.0777852
-0.1090026 -0.0988943 -0.0445145 0.03802977 -0.0484086 -0.0337959
0.07326921 0.02654305 -0.1239398 0.03043288 0.09781751 0.02590732
-0.0586419 -0.08015 -0.0073617 -0.1682889 0.00400978 0.01282504
0.05150735 -0.1449667 0.06144469 0.1005446 0.22570252 -0.3763289
-0.0001517 -0.0521925 0.21106339 -0.4393073 0.0053312 0.13283829
0.12470152 -0.3589714 -0.0061972 0.07370338 0.25447422 -0.3289591
-0.049451 0.05717351 0.14784867 -0.3082401 0.01207511 -0.1141143
0.18880892 -0.3259364 0.04754021 -0.0576587 0.02376083 -0.2828108
0.0234996 -0.1177034 0.02549919 -0.1671077 0.00582423 -0.0715723
0.16712189 -0.0122822 -0.109654 -0.0327367 0.01481733 -0.0636454
-0.0487184 0.01467591 -0.0759871=

0.146753 -0.0931665 -0.1475015 0.07284982 -0.0609536 -0.0945313
-0.0739603 0.17018235 -0.0636651 0.04693379 -0.2586751 0.15550844
-0.1548294 -0.0908961 -0.0415557 0.04915113 -0.0436857 -0.031472
-0.1728483 0.12621336 -0.1321529 -0.1091831 -0.0989133 0.0294641
-0.0950026 -0.1562225 -0.0917397 0.18711324 0.04599057 -0.2039073
0.07691807 0.13016214 0.10801306 -0.3151104 0.0105284 0.10938062
-0.035349 -0.302975 0.03706082 0.12322487 0.07198878 -0.2535323
0.04664604 0.08887579 -0.0210248 -0.1427284 0.09078772 0.08646259
0.00194441 -0.1631221 0.11259725 -0.0984519 -0.0939511 -0.218395
0.13777457 0.00339417 -0.2007502 -0.0703103 0.1548807 0.13540466
-0.0514387 -0.0722146 0.07706029 0.04593663 -0.2334163 -0.0250262

-0.2839164 0.12684187 -0.2450078=

-0.1147067 -0.0084124 -0.5239977 -0.5021591 0.02636886 0.1470097
-0.5139894 -0.6221746 -0.3979228 0.30136263 -0.742976 -0.4011821
0.19038832 0.55414283 -1.1652025 -0.3686967 -0.4750175 0.54713631
-0.9312411 -0.410718 -0.1498093 0.55332947 -1.0870041 -0.4378341
-0.5433689 0.92539561 -0.9013531 -0.6145319 -0.5512772 1.0310978
-0.9422795 -0.6914638 -0.7839714 1.4393494 -0.7092296 -0.894987
-0.6896155 1.1251011 -0.8161536 -0.8204682 -0.8957642 1.3315079
-1.0231192 -0.5556009 -0.7499282 1.281976 -0.9347371 -0.6562014
-0.6568274 1.1967098 -1.150661 -0.5503616 -0.6640182 0.84698498
-0.7811472 -0.5740913 -0.4527726 0.64911795 -0.6970047 -0.5759697
-0.4704399 0.51728982 -0.545236 -0.8311051 -0.4240301 0.37167478
-0.7735854 -0.3031097 -0.4083092 -0.0152683 -0.2330878 -0.5839304
-0.1544528 0.2042688 -0.8989772 -0.3088974 -0.2014994 0.11505035
-0.4815812 -0.5319371 -1.3798244=

0.07143499 -0.1589592 0.04816094 -0.0301291 0.15144217 -0.3037405
0.1549352 -0.0608833 0.21059546 -0.4705076 0.16360784 -0.0684895
0.44703272 -0.6194252 0.19459446 -0.0523894 0.31194624 -0.8030509
0.2595928 -0.119705 0.4913742 -0.8455008 0.15694356 -0.0023983
0.53066176 -0.9705743 0.1324198 0.08982921 0.43900672 -0.8588745
0.1702383 0.02221953 0.44412452 -0.7700244 0.10496679 0.14137991
0.5403164 -0.5077381 0.00849557 0.1611405 0.31764683 -0.5240273
-0.092208 0.21902563 0.25788471 -0.3861519 -0.2022993 0.13711917
0.22238699 -0.156256 -0.2092034 0.16458821 0.20111787 -0.1418906
-0.180493 0.17164391 0.15690604 -0.0254563 -0.1990184 0.10211211
0.17421109 -0.0730809 -0.3717274 0.1436436 -0.0215865 -0.2363243
-0.1982318 0.06996673 0.19735655 0.05625506 -0.241524 0.12768924
0.05979542 -0.0623277 -0.2521037 0.0944353 -0.0492548 0.05238663
-0.1978694 0.05119598 -0.2067173=

0.06230025 -0.0752745 0.32974288 0.00985043 0.07881941 -0.0835249
0.1073643 -0.090154 -0.0938452 0.00704324 0.2569764 0.08700065
-0.0272076 -0.1014201 0.19723812 -0.0935401 0.0913924 -0.0728388
0.33091745 -0.0610701 0.01335303 0.02156818 0.21619918 -0.0909865
0.01069087 0.02569587 0.11676744 -0.0213131 0.1322203 0.11848255
0.11231339 -0.0392407 0.06117272 -0.0234323 0.14693312 0.13509636
-0.0213237 -0.0261696 0.09474246 -0.0100756 0.10580003 -0.0147534
0.12980145 -0.038394 0.08167668 -0.0105376 0.02142166 -0.0161705
0.15833771 0.01835199 0.04420554 0.02605363 0.27427858 0.05774866
-0.0696303 0.03802699 0.0806741 0.03993953 -0.0121658 0.07568218
0.05538817 0.01067943 0.04131892 -0.0267609 0.14418064 0.0897231
-0.0677462 -0.0772208 0.16641215 0.09142463 0.02115551 -0.0876383
0.14652038 0.06084725 -0.1150111 -0.0687876 0.10878915 0.32776353
-0.1929855 0.00694158 0.26604816=

0.1325469 0.15324508 -0.01398 0.08281901 0.07909692 0.36858437
-0.0007111 0.13285491 -0.1658676 0.25348473 0.08835109 0.16466415
-0.118853 0.26435438 -0.0775707 0.09143513 -0.1019902 0.29236633
0.07947435 0.07329605 -0.0903666 0.10754076 0.04456592 0.18368921
-0.162177 0.18712705 0.03216886 0.04698242 -0.0385783 0.2276271
0.04106503 0.08498254 -0.0325038 0.29328787 0.01249749 0.10016124
-0.0012895 0.2371086 0.14713244 -0.053306 -0.0808243 0.28909287
0.13412228 0.10756335 -0.0486093 0.05799349 0.21323961 -0.0118695
-0.142963 0.09792294 0.06907349 0.05942665 -0.143813 0.21673524
0.19903891 0.02989559 0.15750381 -0.0373194 0.12471988 0.10462648
-0.0027455 0.16604523 0.06245366 -0.0775013 -0.0160873 0.21550164
0.25000233 0.05931267 0.22881882=

0.04679342 0.10158926 -0.122116 0.23491009 -0.0625733 0.19985424
-0.1704439 0.302394 -0.0671487 0.33251444 -0.0581705 0.21095584
-0.215752 0.32740423 -0.1597161 0.18950906 -0.1232446 0.27883759
-0.0430407 0.04886867 -0.0914212 0.28192514 0.05275658 0.21014904
-0.1322077 0.2981362 0.1254565 0.15627012 0.04116358 0.08507752
0.10109599 0.23081669 -0.1617257 0.29508773 -0.0405337 -0.0497829
-0.0808031 0.15750171 0.08072432 0.12990661 -0.1935954 0.29120663
0.13912162 0.04256131 -0.1625126 0.25232118 0.04736055 -0.0530935
-0.2270383 0.22945035 0.18167619 0.00080986 -0.1253632 0.15695702
0.01596376 0.03504543 0.00964208 0.11757879 -0.0230768 0.04350457
-0.1284984 0.24145114 0.20540115 0.07580803 -0.0932236 0.14288881
0.00538179 0.05302088 -0.1001294 0.27505419 0.22654785 0.02395938
-0.0861699 0.05814215 0.21307872 0.01372274 0.04515802 -0.0269269
0.20031671 0.23140682 0.16010799=

0.37838998 0.00934576 -0.139213 0.29823828 0.40640026 -0.067578
-0.038453 0.24550894 0.30729383 -0.2807365 -0.0689575 0.26537073
0.58336282 -0.2145292 -0.2378269 0.25939462 0.64761585 -0.3581158
0.07741276 0.45081589 0.65251595 -0.4543131 -0.0671543 0.48592216
0.85640681 -0.6068144 -0.1187844 0.35959438 0.71842372 -0.7140775
-0.0642752 0.37914035 0.71409059 -0.7180941 0.21169594 0.27888221
0.79736245 -0.7102081 0.14268413 0.41374633 0.75569016 -0.7394939
0.02592243 0.37013471 0.82774776 -0.8136597 0.24068722 0.45081198
0.88004726 -0.6990998 0.23456772 0.24596012 0.67229778 -0.8148533
0.30492786 0.39735735 0.55497372 -0.6593497 0.20656242 0.3752968
0.54989374 -0.5660355 0.1205707 0.22377795 0.46045718 -0.519361
0.17151839 0.39539635 0.50465524 -0.3791285 0.07184427 0.36315975
0.51068121 -0.3502096 -0.2094818 0.31471297 0.18174268 -0.1241962
-0.1255455 0.35898197 0.79502285=

0.02952595 -0.0751979 -0.2556099 -0.3040917 -0.0942183 -0.0541431
-0.6262965 -0.1423945 -0.0537339 0.11189342 -0.3791296 -0.3382006
0.02978903 0.20563391 -0.5457558 -0.3666513 -0.1922515 0.29512301

Table 5. Code for running the neural net (neural_n.c).

```
#define local far
#include <windows.h>
#include <alloc.h>
#include "utils.h"
#include <string.h>
#include <ctype.h>
#include <stdio.h>
#include <math.h>
#include <mem.h>
#include "des_util.h"
#include "chipwin.h"
#include "lin_alg.h"

void reportProblem( char local * message, short errorClass);
char iniFileName[] = "designer.ini";

static void sigmoid( vector local * transformMe ){
        short i;
        for( i = 0; i < transformMe->size; i++ )
                transformMe->values[i] =  1/(1+ exp(-1 * transformMe->values[i]));
}

static short getNumCols(char far * buffer){
        short count = 1;
        for( ;*buffer != 0; buffer++ )
                if( *buffer == '\t') count++;
        return count;
}

static short getNumRows(char far * buffer){
        char far * last, far * current;
        short count = -1;
        current = buffer;
        do{
                count++;
                last = current;
                current = strchr( last+1, 0 );
        }while( current > last+1 );
        return count;
}

static void readMatrix( matrix local * theMat, char far * buffer ){
        short i,j;
        char far * temp;
        temp = buffer;
```

```
}

short runForward( vector local *input, vector local *output,
                                                matrix local *weights1, matrix local
*weights2){
        vector hiddenLayer;
        if( !allocateVector( &hiddenLayer, (short)(weights1->numRows +1) )) return
FALSE;
        if( ! vectorTimesMatrix( input, &hiddenLayer, weights1 ) ){
          freeVector( &hiddenLayer ); return FALSE;
        }
        sigmoid( &hiddenLayer );
        hiddenLayer.values[ hiddenLayer.size -1] = 1;
        if( !vectorTimesMatrix( &hiddenLayer, output, weights2 ) ){
          freeVector( &hiddenLayer ); return FALSE;
        }
   freeVector( &hiddenLayer );
        sigmoid( output );
        return TRUE;
}

static vector inputVector= {NULL, 0}, outputVector = {NULL, 0}; static matrix
firstWeights = {NULL, 0, 0} , secondWeights = {NULL, 0, 0};

static short beenHereDoneThis = FALSE;

static short makeSureNetIsSetUp( void ){
        if( beenHereDoneThis ) return TRUE;
        if( !readNeuralNetWeights( &firstWeights, &secondWeights )) return = FALSE;
        if( !allocateVector( &inputVector, firstWeights.numCols )) return = FALSE;
        if( !allocateVector( &outputVector, secondWeights.numRows )) return = FALSE;

        beenHereDoneThis = TRUE;
        return TRUE;
}

void removeNetFromMemory( void ) {
        freeVector( &inputVector ); freeVector( &outputVector );
        freeMatrix( &firstWeights ); freeMatrix( &secondWeights );
   beenHereDoneThis = FALSE;
}

short nnEstimateHybAndXHyb( float local * hyb, float local * xHyb, char = local * probe){
        short probeLength, i;

        if( !makeSureNetIsSetUp()) return FALSE;
        probeLength = (short)(strlen( probe ));
```

```
        return TRUE;
}
void freeVector( vector local * theVec ){
        free( theVec->values );
   theVec->values = NULL;
   theVec->size = 0;
}

void freeMatrix(  matrix local * theMat){
        short i;
        for( i = 0; i < theMat->numRows; i++ )
      free( theMat->values[i] );
        free( theMat->values );
   theMat->values = NULL;
   theMat->numRows = theMat->numCols = 0;
}

float vDot( float local * input1, float local * input2, short size ){
        float returnValue = 0;
        short i;
        for( i = 0; i < size; i++)
                returnValue += input1[i] * input2[i];
        return returnValue;
}

short vectorTimesMatrix( vector local *input, vector local *output,
                                                matrix local *mat ){
        short i;
        if( (input->size != mat->numCols) || (output->size < mat->numRows) ){
                errorHwnd( "illegal multiply" );
                return FALSE;
        }
        for( i = 0; i < mat->numRows; i++ )
                output->values[i] = vDot( input->values, mat->values[i], input->size =
);
        return TRUE;
}
```

It is understood that the examples and embodiments described herein are for illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the appended claims. All publications, patents, and patent applications cited herein are hereby incorporated by reference for all purposes.

5. The method of claim 1, wherein said oligonucleotides are from 5 to 45 nucleotides in length

6. The method of claim 7, wherein said oligonucleotides are from 20 to 25 nucleotides in length

7. The method of claim 1, wherein said oligonucleotides are synthesized by light-directed polymer synthesis.

8. The method of claim 1, wherein said array comprises oligonucleotide sequences from constitutively expressed control genes

9. The method of claim 8, wherein said control genes are selected from the group consisting of ß-actin, GAPDH, and the transferrin receptor.

10. The method of claim 1, wherein the variation between different copies of each array is less than 20% wherein said variation is measured as the coefficient of variation in hybridization intensity averaged over at least 5 oligonucleotide probes for each gene whose expression the array is to detect.

11. The method of claim 1, wherein said pool of target nucleic acids is labeled with a single species of fluorophore.

12. The method of claim 1, wherein preparation of said oligonucleotide probes does not require cloning, a nucleic acid amplification step, or enzymatic synthesis.

13. The method of claim 1, wherein preparation of said oligonucleotide probes does not require handling of any biological materials.

14. The method of claim 1, wherein the concentration of nucleic acids in said pool is proportional to the expression levels of said genes.

24. The method of claim 1, wherein said hybridization is performed with a fluid volume of about 250 μl or less.

25. The method of claim 1, wherein said quantifying comprises detecting a hybridization signal that is proportional to the concentration of said RNA in said nucleic acid sample.

26. The method of claim 1, wherein said quantifying comprises detecting a hybridization signal that is proportional to the concentration of said target nucleic acids for each gene in said pool of target nucleic acids.

27. The method of claim 1, wherein said hybridization comprises a hybridization at low stringency of 30°C to 50°C and 6 X SSPE-T or lower and a wash at higher stringency.

28. The method of claim 1, wherein said pool of nucleic acids is a pool of mRNAs.

29. The method of claim 1, wherein said pool of nucleic acids is a pool of RNAs *in vitro* transcribed from a pool of cDNAs.

30. The method of claim 1, wherein said pool of nucleic acids is amplified from a biological sample.

31. The method of claim 1, wherein said pool of nucleic acids comprises fluorescently labeled nucleic acids.

32. The method of claim 1, wherein said detecting comprises quantifying fluorescence of a label on said hybridized nucleic acids at a spatial resolution of 100 μm or higher.

(c) isolating or amplifying the remaining polyA$^+$ RNA in said pool.

37. A composition indication the expression levels of a multiplicity of genes, said composition comprising an array of oligonucleotide probes immobilized on a surface, said array comprising more than 100 different oligonucleotides wherein:

each different oligonucleotide is localized in a predetermined region of said surface;

each different oligonucleotide is attached to said surface through a single covalent bond;

the density of said different oligonucleotides is greater than about 60 different oligonucleotides per 1 cm^2; and

said oligonucleotide probes are complementary to subsequences of said genes; and

said oligonucleotide probes are specifically hybridized to one or more fluorescently labeled nucleic acids forming a fluorescent array such that the fluorescence of said array is indicative of the transcription levels of said multiplicity of genes.

38. The composition of claim wherein said fluorescence intensity is proportional to the transcription levels of said multiplicity of preselected genes in a biological sample.

39. The composition of claim wherein said array of oligonucleotides further comprises mismatch control probes.

40. The method of claim 37, wherein each of said oligonucleotide probes is chemically synthesized.

41. The composition of claim 40, wherein said oligonucleotides are from 5 to 45 nucleotides in length.

where, for each gene of said multiplicity of genes, said array includes at least one oligonucleotide probe complementary to a subsequence of said gene; and

instructions describing the use of said array for the quantification of expression levels of said multiplicity of genes.

49. The kit of claim 48, wherein said oligonucleotide probes range in length from 5 to 45 nucleotides.

50. The kit of claim 48, wherein said array further comprises mismatch control probes such that for each probe specific to a gene there exists a mismatch control probe.

51. The kit of claim 48, further comprising fluorescent label for labeling RNA or DNA that is to be hybridized to the oligonucleotides of said array.

52. The kit of claim 48, further comprising buffers and reagents for the hybridization of RNA to the oligonucleotide probes of said array.

53. A method of selecting a set of oligonucleotide probes that specifically bind to one or more target nucleic acids, said method comprising:

(a) providing a high density array of oligonucleotide probes said array comprising a multiplicity of oligonucleotide probes, wherein each probe is complementary to a subsequence of said target nucleic acids and for each probe there is a corresponding mismatch control probe;

(b) hybridizing said target nucleic acids to said array of oligonucleotide probes; and

(c) selecting those probes where the difference in hybridization signal intensity between each probe and its mismatch control is detectable.

54. The method of claim 53, further comprising:

62. The method of claim 53, wherein said mismatch control probes have a centrally located 1 base mismatch.

63. The method of claim 53, wherein said hybridization comprises hybridization at low stringency of 30°C to 50°C and 6 X SSPE-T or lower followed by one or more washes at progressively increasing stringency until a desired level of hybridization specificity is obtained.

64. The method of claim 63, wherein pool of nucleic acids is a pool of nucleic acids having a sense opposite that of the nucleic acids to which said oligonucleotide probes are complementary.

65. In a computer system, a method of monitoring expression of genes, the method comprising the steps of:

receiving input of hybridization intensities for a plurality of nucleic acid probes including pairs of perfect match probes and mismatch probes, the hybridization intensities indicating hybridization affinity between the plurality of nucleic acid probes and nucleic acids corresponding to a gene, and each pair including a perfect match probe that is perfectly complementary to a portion of the nucleic acids and a mismatch probe that differs from the perfect match probe by at least one nucleotide;

comparing the hybridization intensities of the perfect match and mismatch probes of each pair; and

indicating expression of the gene according to results of the comparing step.

66. The method of claim 65, wherein the comparing step includes the step of calculating differences between the hybridization intensities of the perfect match and mismatch probes of each pair.

67. The method of claim 66, wherein the comparing step includes the step of calculating an average of the differences.

75. The method of claim 74, wherein the identifying step includes the step of analyzing each probe of the set by criteria that specify characteristics indicative of low hybridization or high cross hybridization.

76. The method of claim 75, wherein each of the criteria includes a threshold value such that if a selected probe has a characteristic that crosses the threshold value, low hybridization or high cross hybridization are indicated for the selected probe.

77. The method of claim 76, further comprising the step of increasing at least one threshold value to increase the probes in the subset.

78. The method of claim 75, wherein the identifying step is performed by a neural network that receives as input the probes of the set and outputs the probes of the subset.

79. The method of claim 75, further comprising the step of determining the criteria as heuristic rules derived from multiple experiments.

80. The method of claim 75, wherein one of the criteria indicates low hybridization or cross hybridization if occurrences of a specific nucleotide in a probe crosses a threshold value.

81. The method of claim 75, wherein one of the criteria indicates low hybridization or cross hybridization if a number of a specific nucleotide that repeats sequentially in a probe crosses a threshold value.

82. The method of claim 75, wherein one of the criteria indicates low hybridization or cross hybridization if a length of a palindrome in a probe crosses a threshold value.

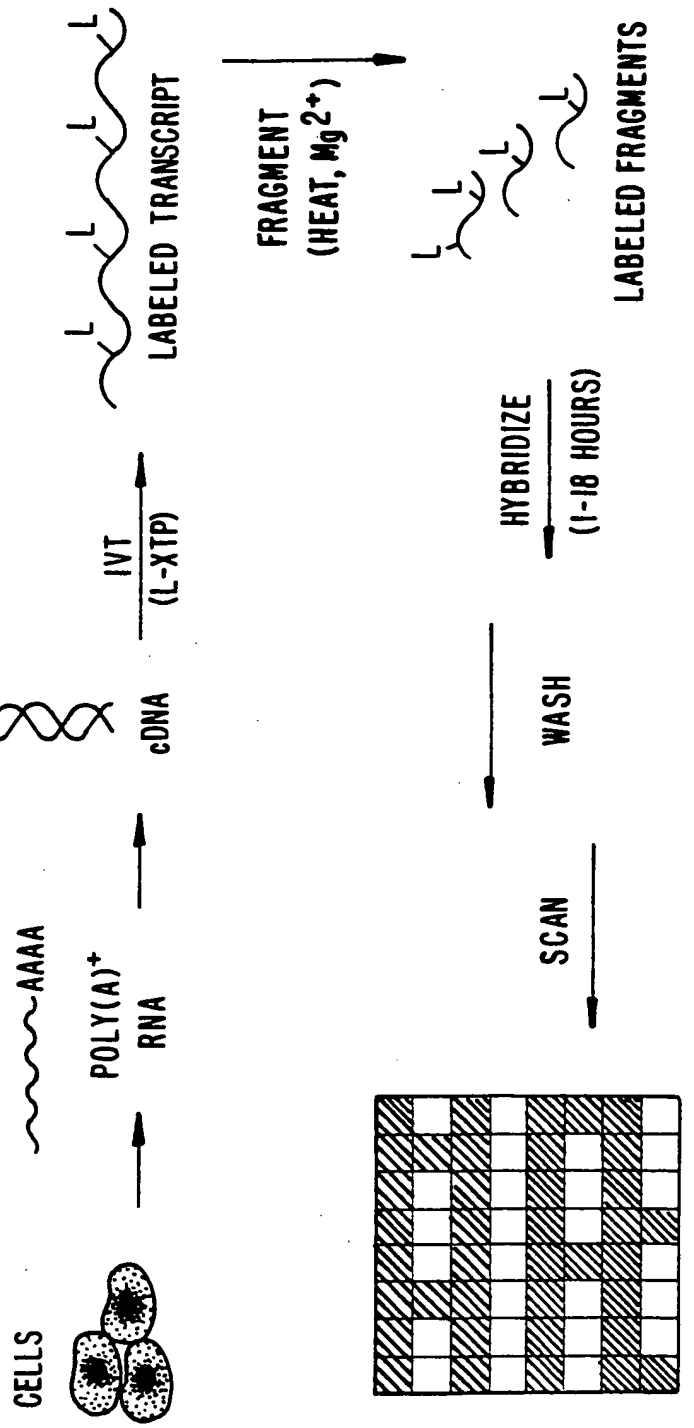

CELLS
POLY(A)+ RNA
AAAA
cDNA
IVT (L-XTP)
LABELED TRANSCRIPT
FRAGMENT (HEAT, Mg2+)
LABELED FRAGMENTS
HYBRIDIZE (1-18 HOURS)
WASH
SCAN

FIG. 1.

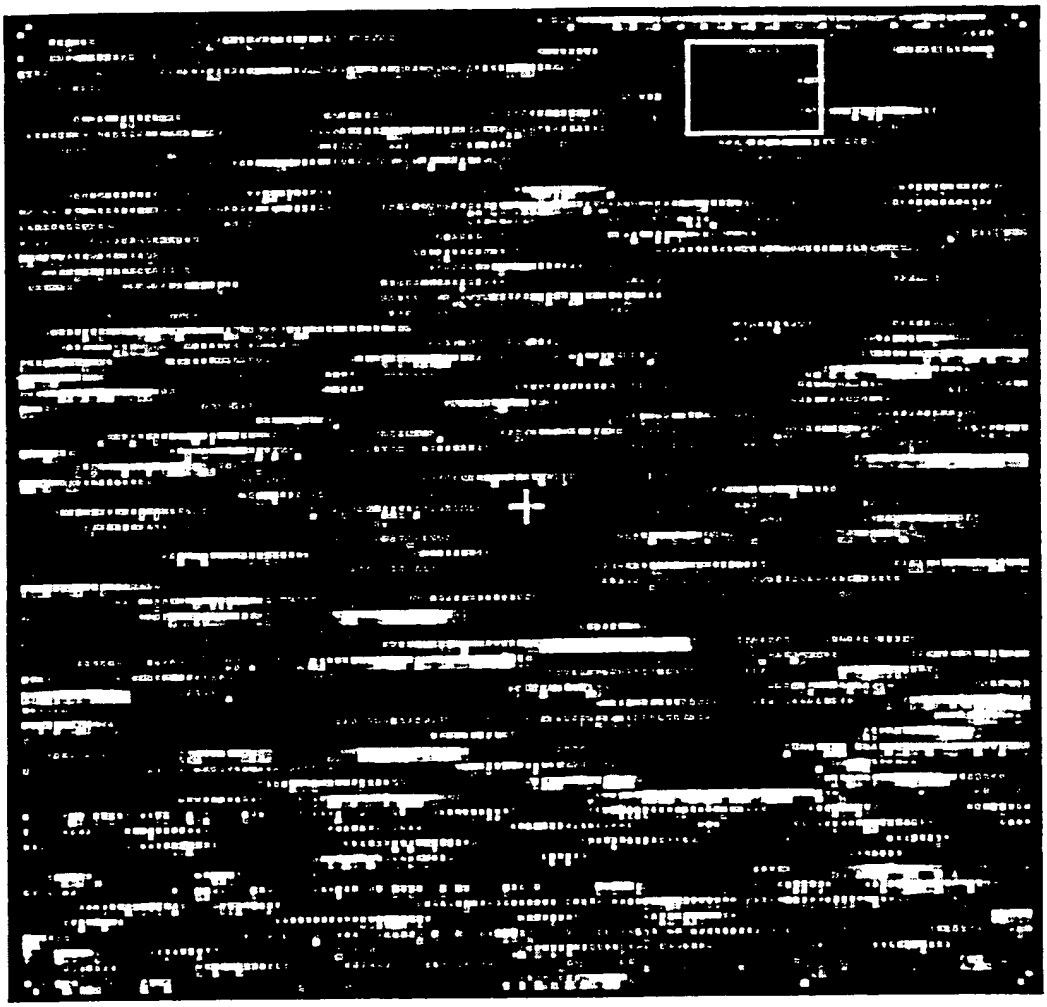


FIG. 2A.

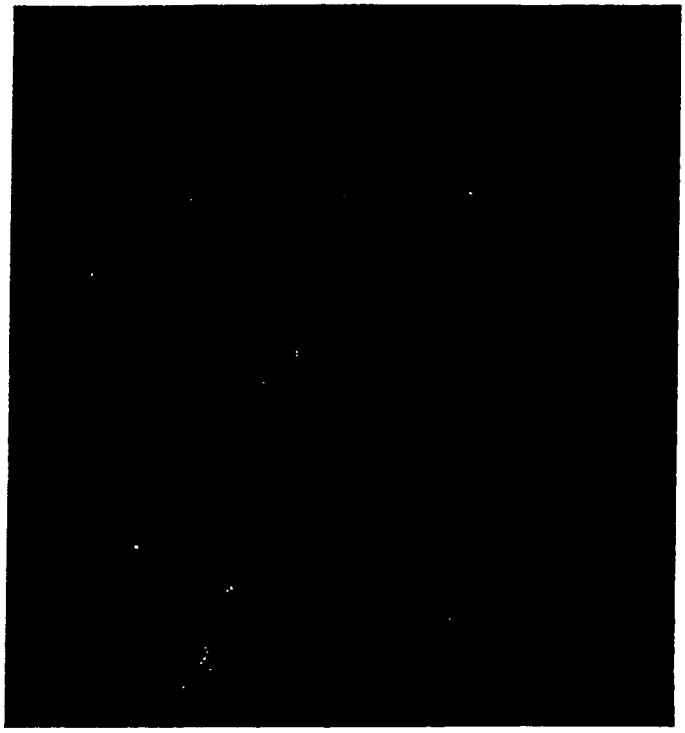


FIG. 2C.

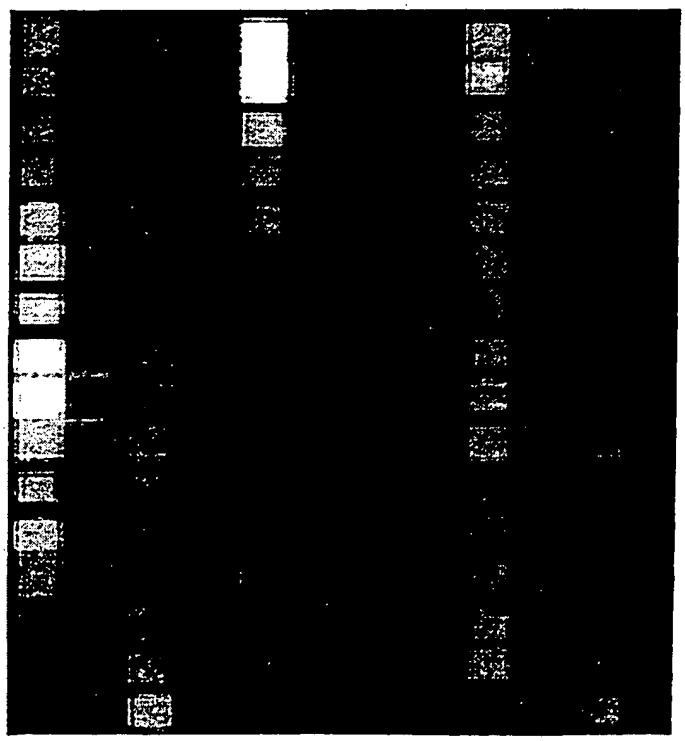


FIG. 2B.

2M 1M

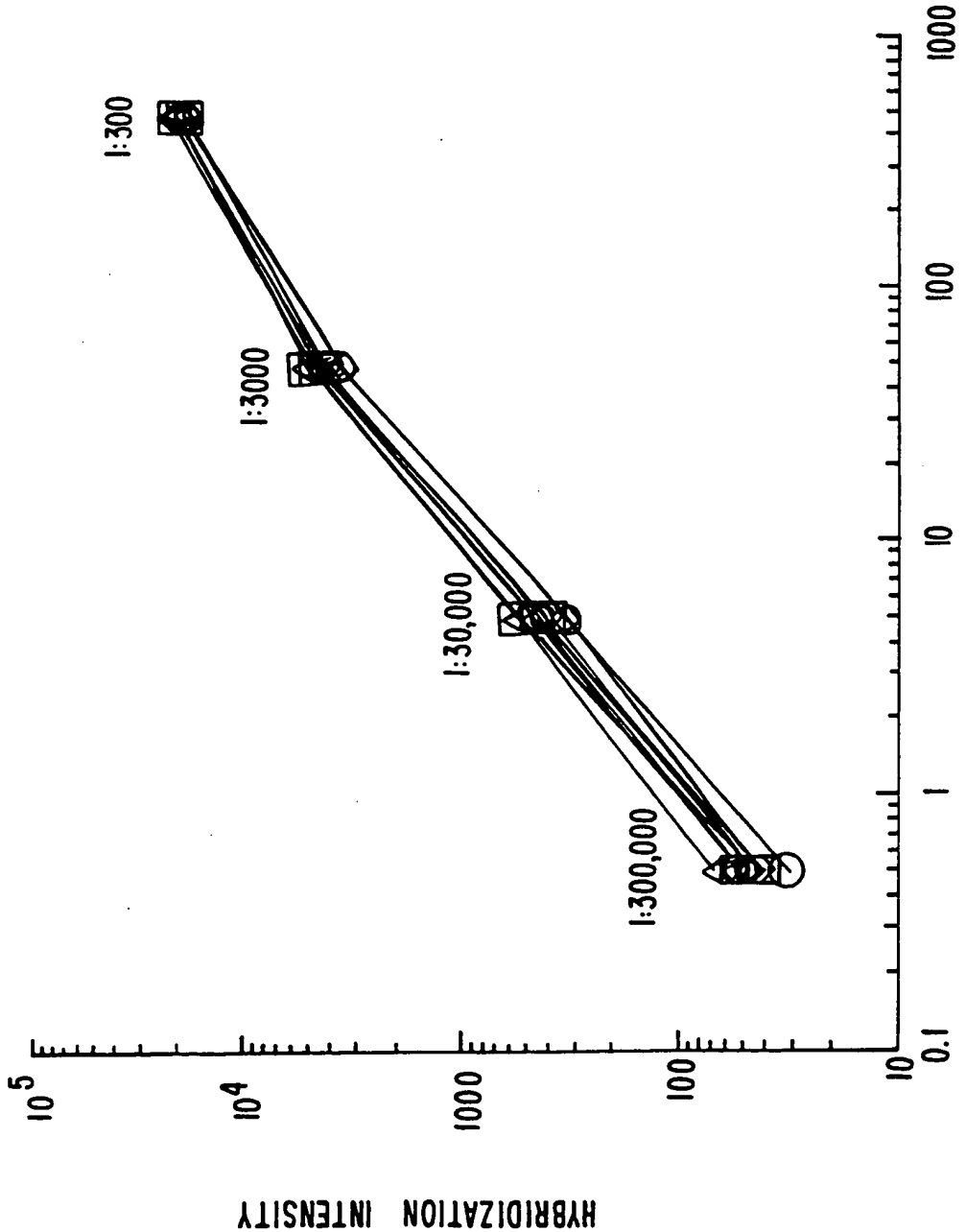

HYBRIDIZATION INTENSITY
10^5
10^4
1000
100
10
0.1
1
10
100
1000
RNA CONCENTRATION (pM)
1:300
1:3000
1:30,000
1:300,000

FIG. 3.

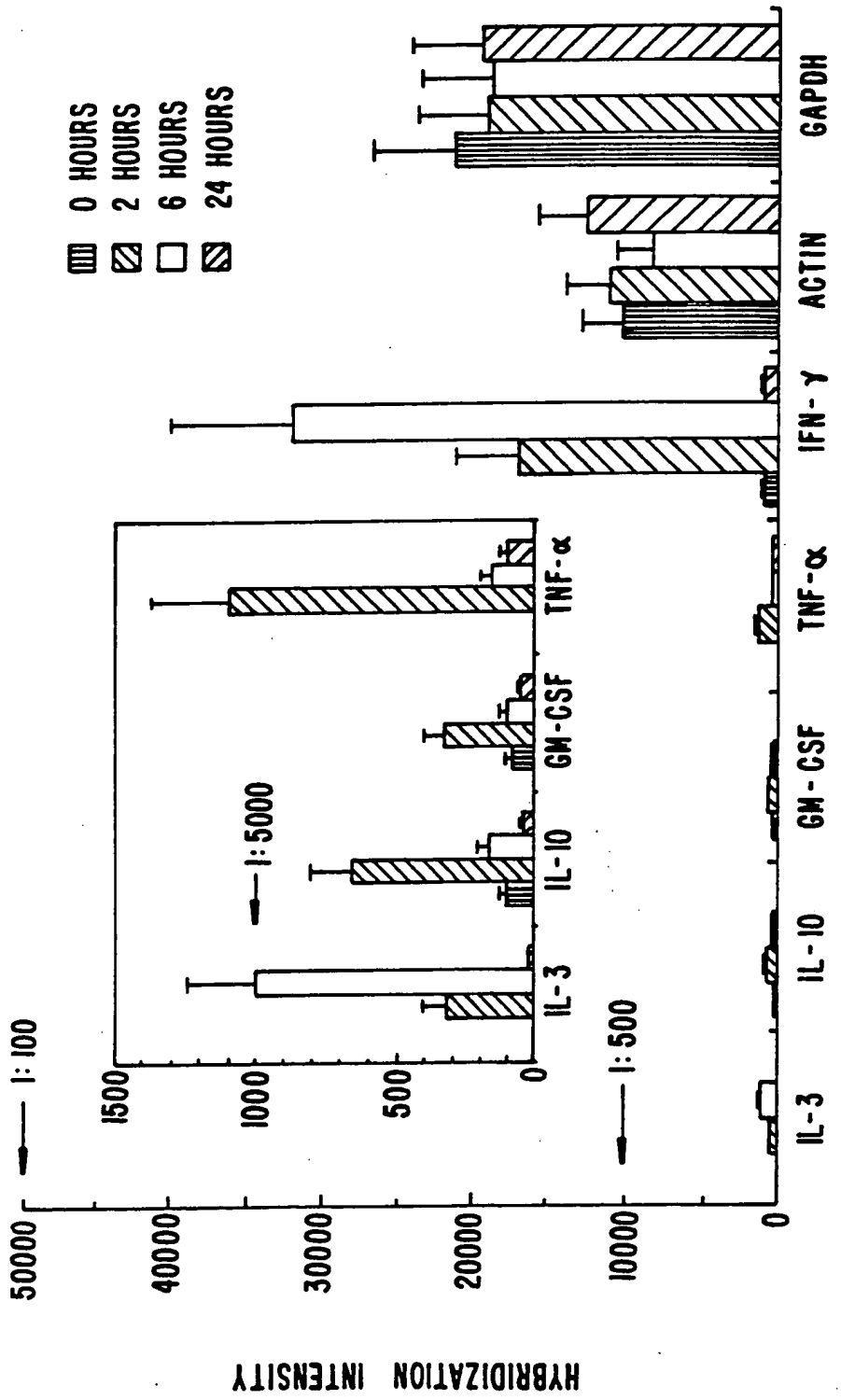

0 HOURS
2 HOURS
6 HOURS
24 HOURS
GAPDH
ACTIN
IFN-γ
TNF-α
GM-CSF
IL-10
IL-3
1:100
1:500
1:5000
50000
40000
30000
20000
10000
0
1500
1000
500
0
HYBRIDIZATION INTENSITY

FIG. 4.

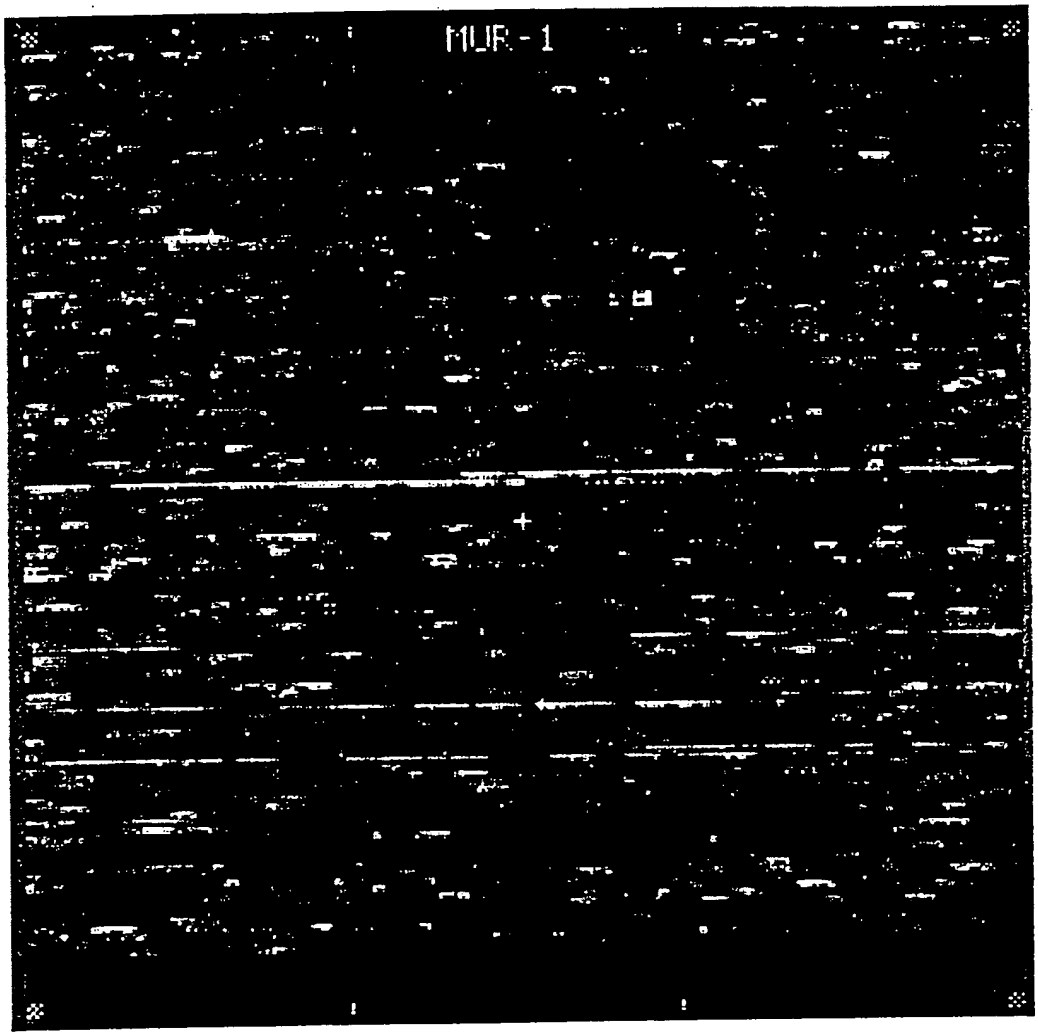


FIG. 5.

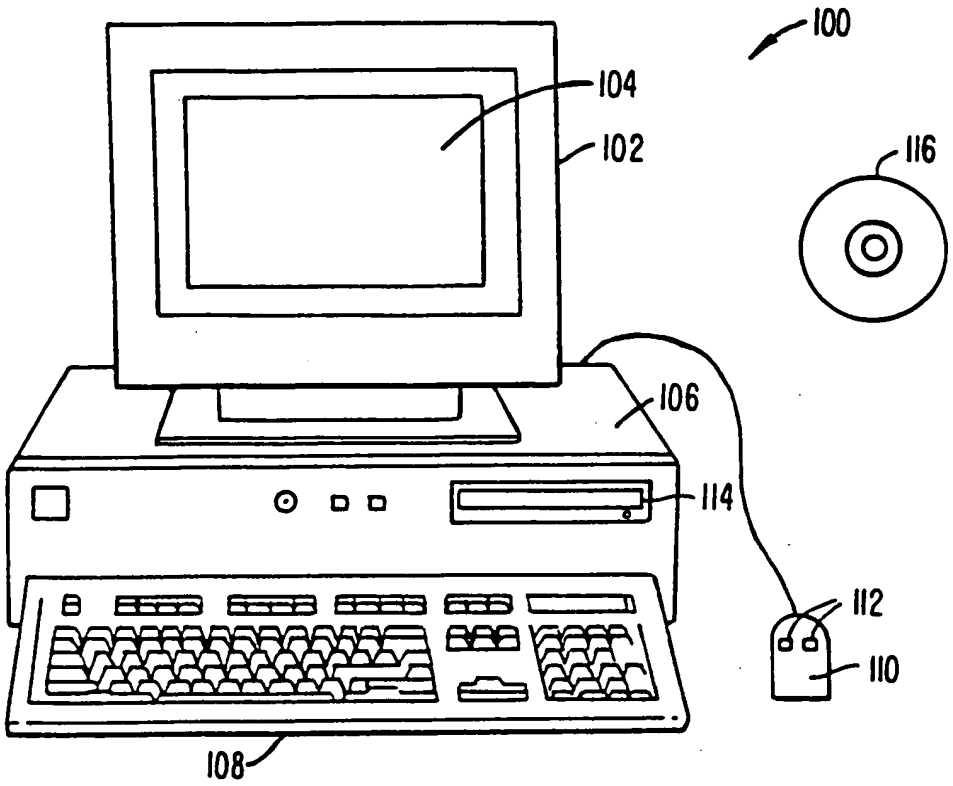

100
104
102
116
106
114
112
110
108

FIG. 6.

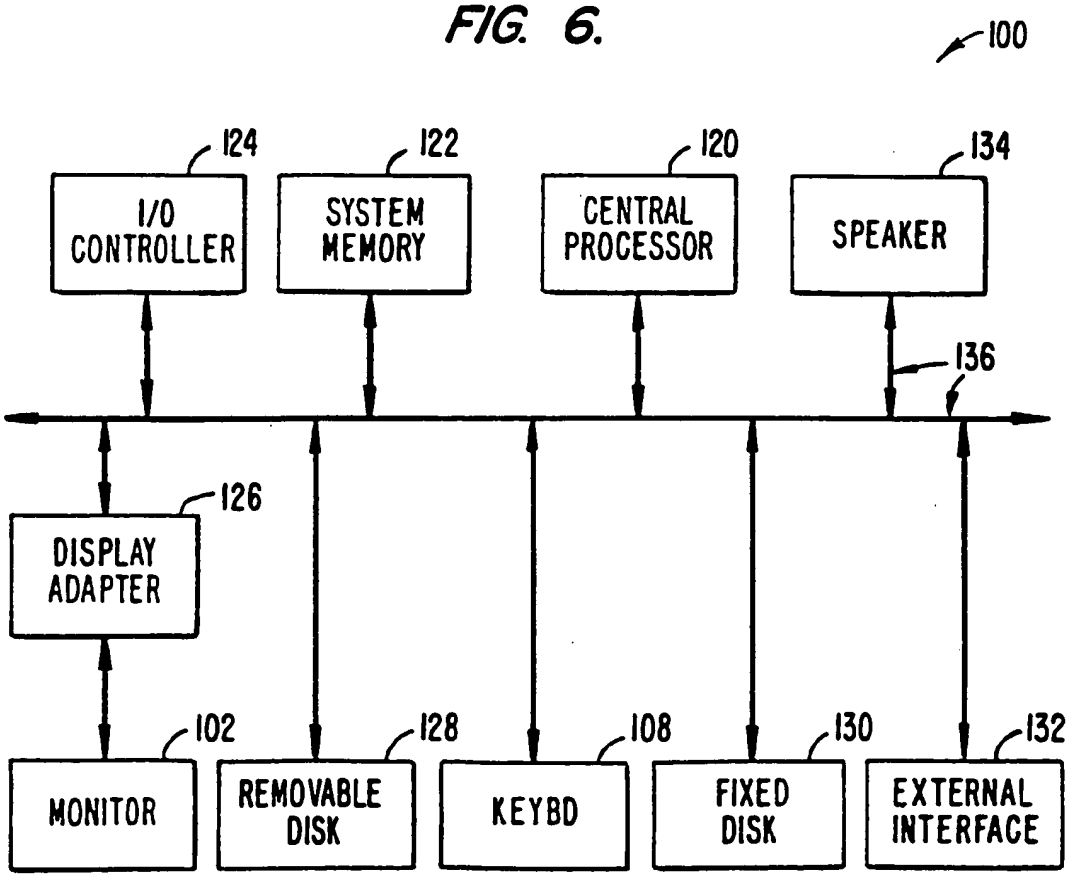

100
124
I/O CONTROLLER
122
SYSTEM MEMORY
120
CENTRAL PROCESSOR
134
SPEAKER
136
126
DISPLAY ADAPTER
102
MONITOR
128
REMOVABLE DISK
108
KEYBD
130
FIXED DISK
132
EXTERNAL INTERFACE

FIG. 7.

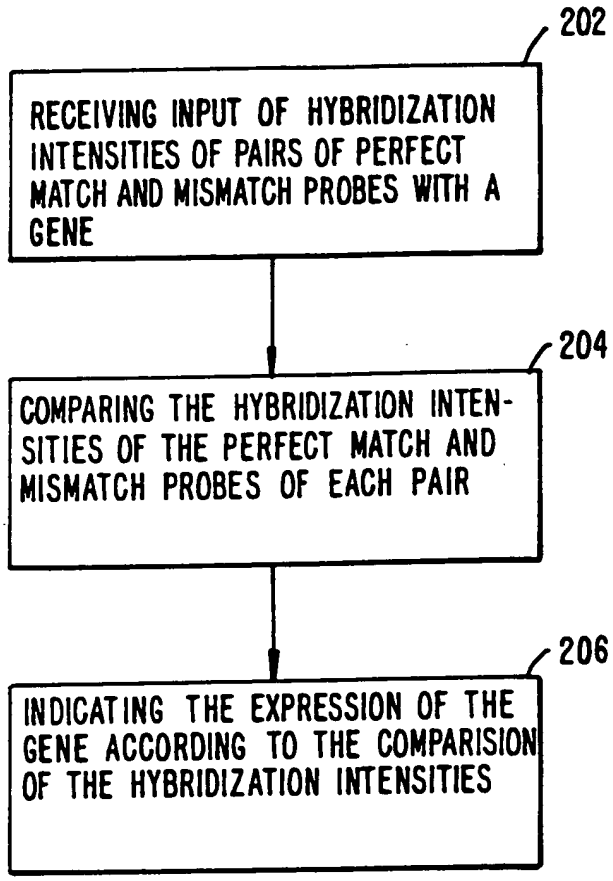

202
RECEIVING INPUT OF HYBRIDIZATION INTENSITIES OF PAIRS OF PERFECT MATCH AND MISMATCH PROBES WITH A GENE
204
COMPARING THE HYBRIDIZATION INTENSITIES OF THE PERFECT MATCH AND MISMATCH PROBES OF EACH PAIR
206
INDICATING THE EXPRESSION OF THE GENE ACCORDING TO THE COMPARISION OF THE HYBRIDIZATION INTENSITIES

FIG. 8.

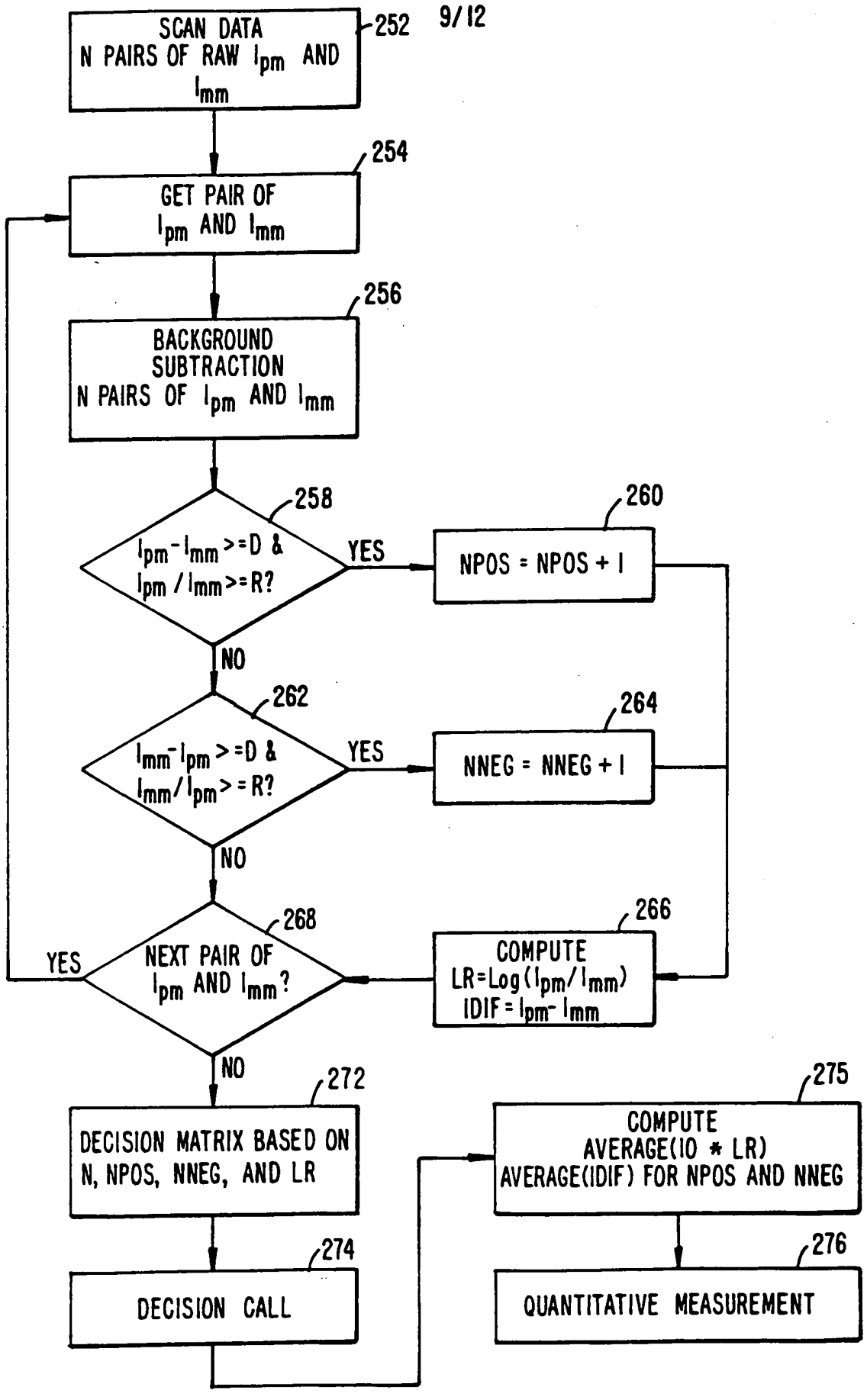

SCAN DATA
N PAIRS OF RAW I_{pm} AND I_{mm}
252
GET PAIR OF I_{pm} AND I_{mm}
254
BACKGROUND SUBTRACTION
N PAIRS OF I_{pm} AND I_{mm}
256
$I_{pm}-I_{mm}>=D$ & $I_{pm}/I_{mm}>=R$?
258
YES
NPOS = NPOS + 1
260
NO
$I_{mm}-I_{pm}>=D$ & $I_{mm}/I_{pm}>=R$?
262
YES
NNEG = NNEG + 1
264
NO
NEXT PAIR OF I_{pm} AND I_{mm}?
268
YES
COMPUTE
LR=Log(I_{pm}/I_{mm})
IDIF = $I_{pm}-I_{mm}$
266
NO
DECISION MATRIX BASED ON N, NPOS, NNEG, AND LR
272
DECISION CALL
274
COMPUTE
AVERAGE(IO * LR)
AVERAGE(IDIF) FOR NPOS AND NNEG
275
QUANTITATIVE MEASUREMENT
276

FIG. 9

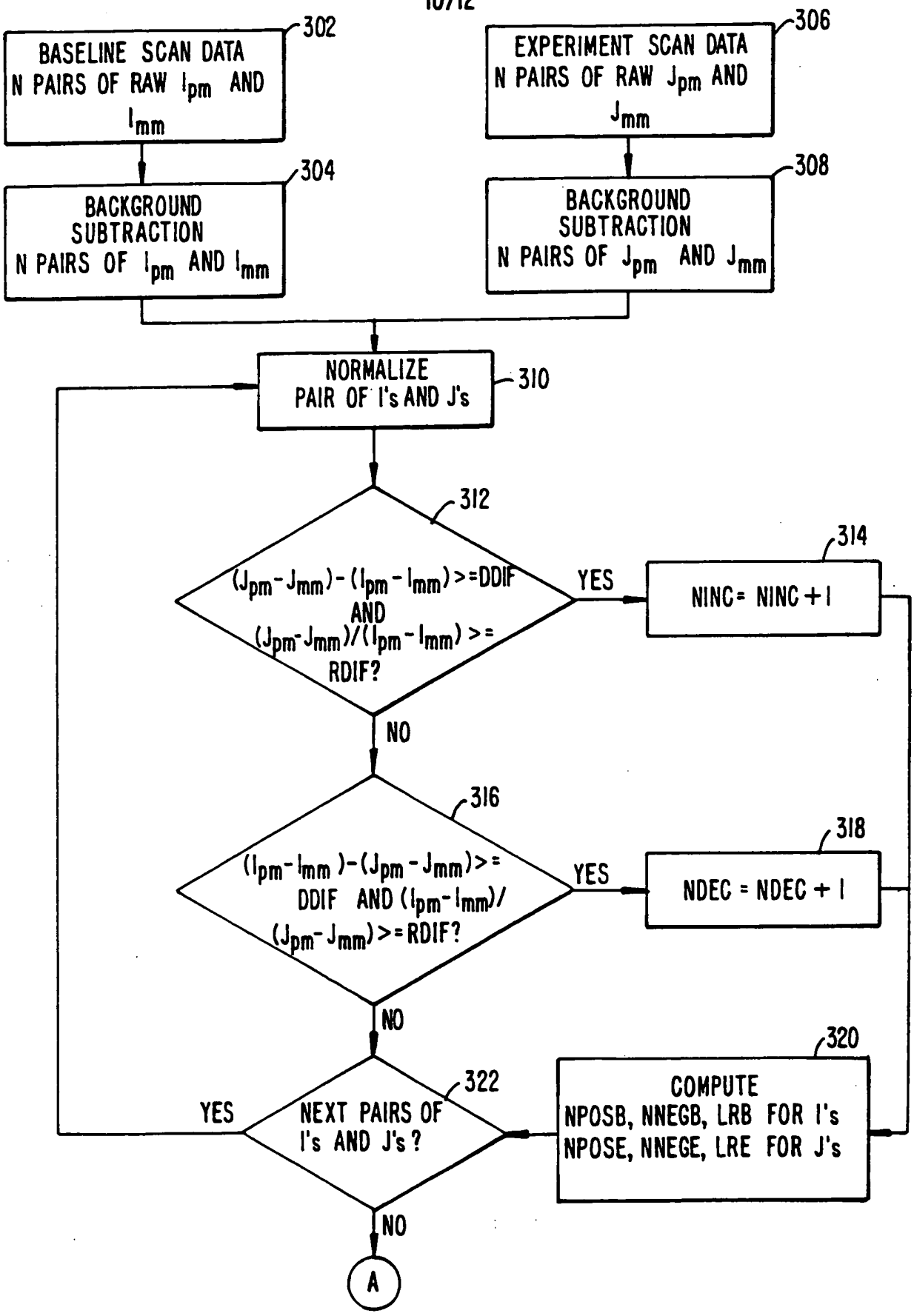

BASELINE SCAN DATA N PAIRS OF RAW I_{pm} AND I_{mm}
302
EXPERIMENT SCAN DATA N PAIRS OF RAW J_{pm} AND J_{mm}
306
BACKGROUND SUBTRACTION N PAIRS OF I_{pm} AND I_{mm}
304
BACKGROUND SUBTRACTION N PAIRS OF J_{pm} AND J_{mm}
308
NORMALIZE PAIR OF I's AND J's
310
$(J_{pm}-J_{mm})-(I_{pm}-I_{mm})>=$DDIF AND $(J_{pm}-J_{mm})/(I_{pm}-I_{mm})>=$ RDIF?
312
YES
NINC= NINC + 1
314
NO
$(I_{pm}-I_{mm})-(J_{pm}-J_{mm})>=$ DDIF AND $(I_{pm}-I_{mm})/(J_{pm}-J_{mm})>=$ RDIF?
316
YES
NDEC = NDEC + 1
318
NO
NEXT PAIRS OF I's AND J's?
322
YES
COMPUTE NPOSB, NNEGB, LRB FOR I's NPOSE, NNEGE, LRE FOR J's
320
NO
A

FIG. 10A

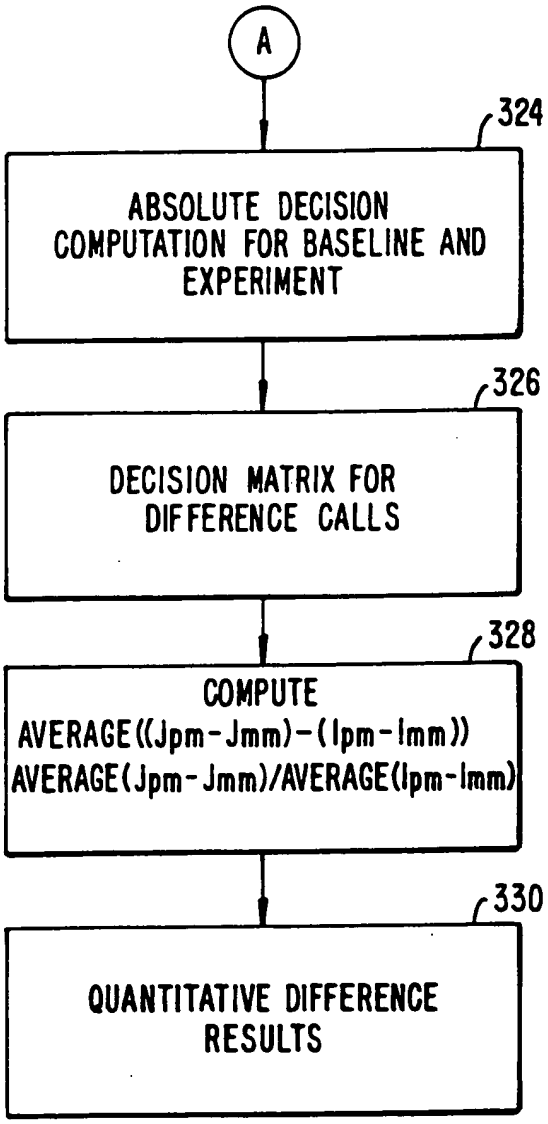
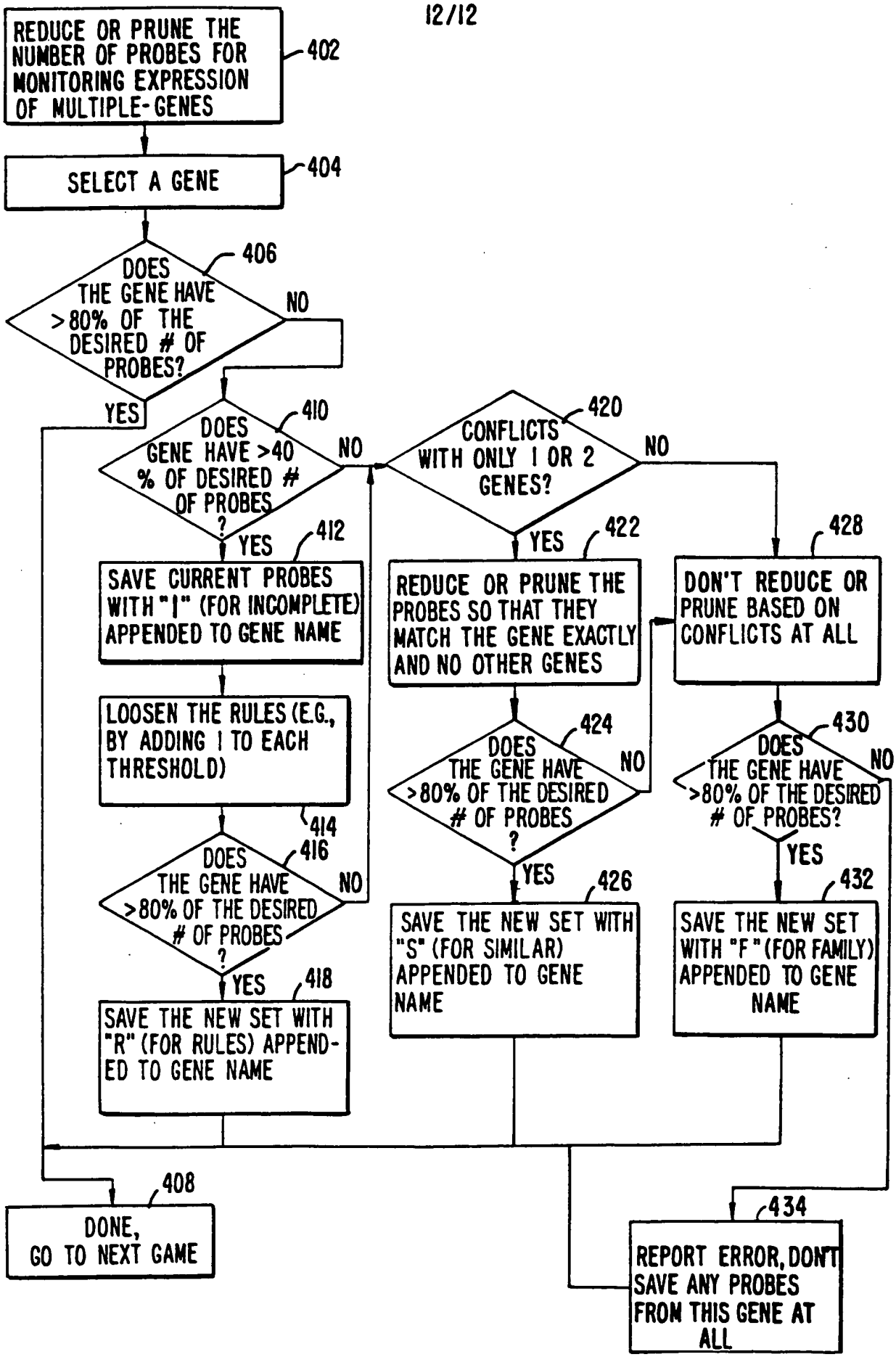

A
324
ABSOLUTE DECISION COMPUTATION FOR BASELINE AND EXPERIMENT
326
DECISION MATRIX FOR DIFFERENCE CALLS
328
COMPUTE
AVERAGE((Jpm-Jmm)-(Ipm-Imm))
AVERAGE(Jpm-Jmm)/AVERAGE(Ipm-Imm)
330
QUANTITATIVE DIFFERENCE RESULTS

FIG. 10B.


REDUCE OR PRUNE THE NUMBER OF PROBES FOR MONITORING EXPRESSION OF MULTIPLE-GENES
402
SELECT A GENE
404
DOES THE GENE HAVE >80% OF THE DESIRED # OF PROBES?
406
NO
YES
DOES GENE HAVE >40 % OF DESIRED # OF PROBES ?
410
NO
YES
SAVE CURRENT PROBES WITH "I" (FOR INCOMPLETE) APPENDED TO GENE NAME
412
LOOSEN THE RULES (E.G., BY ADDING 1 TO EACH THRESHOLD)
414
DOES THE GENE HAVE >80% OF THE DESIRED # OF PROBES ?
416
NO
YES
SAVE THE NEW SET WITH "R" (FOR RULES) APPENDED TO GENE NAME
418
CONFLICTS WITH ONLY 1 OR 2 GENES?
420
NO
YES
REDUCE OR PRUNE THE PROBES SO THAT THEY MATCH THE GENE EXACTLY AND NO OTHER GENES
422
DOES THE GENE HAVE >80% OF THE DESIRED # OF PROBES ?
424
NO
YES
SAVE THE NEW SET WITH "S" (FOR SIMILAR) APPENDED TO GENE NAME
426
DON'T REDUCE OR PRUNE BASED ON CONFLICTS AT ALL
428
DOES THE GENE HAVE >80% OF THE DESIRED # OF PROBES?
430
NO
YES
SAVE THE NEW SET WITH "F" (FOR FAMILY) APPENDED TO GENE NAME
432
DONE, GO TO NEXT GAME
408
REPORT ERROR, DON'T SAVE ANY PROBES FROM THIS GENE AT ALL
434